2018

Characterization of CRISPR RNA guided immunity in Bacillus halodurans type I-C system

Hayun Lee
Iowa State University

Characterization of CRISPR RNA guided immunity in *Bacillus halodurans* type I-C system

by

Hayun Lee

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Biochemistry

Program of Study Committee:
Dipali Sashital, Major Professor
Richard Honzatko
Scott Nelson
Amy Andreotti
Vincenzo Venditti

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2018

DEDICATION

This thesis is dedicated to my dearest parents: mom, Yunchung, and dad, Chungha.

TABLE OF CONTENTS

Page

ACKNOWLEDGMENTS v

ABSTRACT vi

CHAPTER 1. INTRODUCTION: INSIGHTS INTO CRISPR-CAS ADAPATIVE IMMUNE SYSTEM 1
- Overview of CRISPR-Cas Systems and Their Diversity 3
  - Type I-C CRISPR-Cas systems in *Bacillus halodurans* 6
- Adaptation 8
  - Generation of prespacers 10
  - Prespacer selection and processing 13
  - Integration into CRISPR array 16
- CRISPR RNA biogenesis and Cascade formation 19
- Target Binding and Interference 22
- Type I-C CRISPR-Cas systems 26
- Organization of the dissertation 29
- References 30

CHAPTER 2. CAS4-DEPENDENT PRESPACER PROCESSING ENSURES HIGH-FIDELITY PROGRAMMING OF CRISPR ARRAYS 42
- Abstract 42
- Introduction 42
- Results 45
  - Complex formation by type I-C Cas4, Cas1 and Cas2 45
  - Molecular architecture of the Cas4-Cas1 complex 47
  - Cas4 enhances prespacer processing 49
  - PAM-dependent prespacer processing 53
  - Cas4 ensures the integration of processed prespacers 54
  - Sequence-specific integration and asymmetric prespacer processing by the adaptation complex 56
- Discussion 62
- Materials and Methods 66
- References 78

CHAPTER 3. CAS4-CAS1-CAS2 CRISPR ADAPTATION COMPLEX PROCESSES SINGLE STRAND DNA SEQUENCE SPECIFICALLY 82
- Introduction 82
- Results 84
  - Formation of the Cas4-Cas1-Cas2 complex 84
  - Architecture of the Cas4-Cas1-Cas2 complex 85
  - Cas4 is activated for ssDNA processing in the presence of dsDNA 88
  - Precise PAM-specific ssDNA processing by Cas4-Cas1-Cas2 92
- Discussion 95

Materials and Methods 100
References 106

CHAPTER 4. DNA TARGETING BY TYPE I-C CASCADE AND CONSTRUCTION OF MINIMAL-CYSTEINE VARIANT FOR SMFRET 110
Introduction 110
Results 113
Construction of type I-C Cascade for binding assay 113
Cascade/I-C binds non-specific targets 114
Construction of smFRET assay for Cascade-target binding 116
Discussion 118
Materials and Methods 120
References 123

CHAPTER 5. CONCLUSIONS AND FUTURE DIRECTIONS 127
Conclusion 127
Future Directions 129
References 131

ACKNOWLEDGMENTS

My journey towards to this thesis would have not been possible without people who gave me all the encouragement, inspiration, and motivation. For the foremost, I thank my major professor, Prof. Dipali Sashital, for the continuous supports, guidance and patience throughout my study. She groomed me how to be a sound professional and guided me on the right path. I could not have imagined having a better advisor and mentor like her. I also thank my committee members for their critical comments and insightful suggestions to my research projects: Prof. Richard Honzatko, Prof. Scott Nelson, Prof. Amy Andreotti, and Prof. Vincenzo Venditti.

I would like to thank former and current members of the Sashital lab for exploring the CRISPR world with me: Chaoyou, John, Karthik, Phong, Shravanti, and Michael. Many thanks to faculties, staffs, now and then Ph.D students at the BBMB department. You all have made my life in lab pleasant and unforgettable. Also, I extend my acknowledgements to my friends outside the lab for sharing love, friendships, and aspirations with me.

Finally, I am grateful to my family for continually supporting me throughout my life. Chungha Lee and Yunchung Kim, my dad and my mom, have always cheered me on pursuing my dreams and believed in me that I can do this! All your love and supports always gave me power to move one step forward! I also thank my brother, Changin Lee, for supporting me spiritually.

ABSTRACT

Prokaryotes utilize the CRISPR (Clustered Regularly Interspaced Short Palindromic Repeats) – Cas (CRISPR-associated) adaptive immune system to defend against infection. A CRISPR locus consists of an AT-rich leader region followed by a series of DNA repeats interspersed by foreign DNA-derived spacers. Upon viral infection, Cas proteins acquire short fragments from the invader and insert them as new spacers into the CRISPR locus. CRISPR transcripts are generated from the CRISPR locus and assemble with Cas proteins to form the surveillance complex. The CRISPR RNA guides the complex to target foreign genetic elements bearing sequence complementarity to the crRNA and recruits a Cas nuclease for degradation. The research presented in this dissertation focuses on understanding the mechanisms of CRISPR RNA guided immunity in *Bacillus halodurans* type I-C system during adaptation and interference.

Cas4 is widespread across types I, II and V and is thought to be involved in spacer acquisition along with the universally conserved Cas1 and Cas2 proteins, but the role of Cas4 has remained unclear. Using a combination of biochemical and structural experiments, we reveal that type I-C Cas4 in *B. halodurans* interacts directly with Cas1 and Cas2, forming a Cas4-Cas1-Cas2 complex, that mediates spacer selection, processing, and integration during CRISPR immunity. Cas4 associates tightly with Cas1 and the presence of CRISPR DNA substrates helps to stabilize the higher order complex. Cas4 selectively captures spacers that contain protospacer adjacent motifs (PAMs), short sequences required for proper target recognition by the surveillance complex, and processes the substrate directly upstream of the PAM site. When in complex with Cas1-Cas2, Cas4 cleaves spacers endonucleolytically and the complex preferentially integrates the processed spacers at the leader-repeat junction in the CRISPR locus. Together, our

findings demonstrate that Cas4 is indispensable in CRISPR immunity by providing functional spacers for target recognition.

For target recognition, type I-C system is unique in that only three proteins are required to form its surveillance complex. It is unknown how type I-C Cascade searches for targets using this minimal machinery. We investigated binding interactions of *B. halodurans* type I-C Cascade with dsDNA and found that, unlike *E. coli* type I-E Cascade, type I-C Cascade has much strong non-specific affinity for DNA. These observations suggest a search mechanism involving longer-lived interactions with DNA, potentially through one-dimensional sliding. To test this, we initiated development of a single-molecule fluorescence resonance energy transfer (FRET) assay to directly visualize how Cascade searches target DNA in real time. We constructed a system suitable for labeling type I-C Cascade with a fluorophore for the smFRET assay. Using this system, we detected bulk FRET between Cy3-labelled dsDNA target and Cy5-labelled Cascade upon DNA binding. These experiments established a FRET system that will be used for future smFRET experiments to understand the kinetics and mechanisms for searching DNA targets by type I-C Cascade.

CHAPTER 1. INTRODUCTION: INSIGHTS INTO CRISPR-CAS ADAPATIVE IMMUNE SYSTEM

Prokaryotic phages are the most abundant life forms and one of the planet's oldest predators (Breitbart and Rohwer, 2005). Their abundance – outnumbering microbial cells by 10-fold – and higher degree of genomic variability have a critical impact on microbial communities (Chibani-chennoufi et al., 2004; Hatfull, 2008). Consequently, this rapidly evolving and diverse challenge has led to the development of natural defense mechanisms in bacteria that target each step of phage life cycle (Labrie et al., 2010). The first line of defense includes physical or chemical barriers on cell surface receptors, disrupting phage absorption (Forde and Fitzgerald, 1999; Dy et al., 2014). Once attached to suitable receptors, superinfection exclusion systems can block phage DNA injection to the host cell (Seed, 2015). However, upon successful entry into bacterial cells, phage DNA is subject to the well-characterized restriction-modification systems. These systems rely on modification of host DNA by methylation to discriminate self (host) vs. non-self (phage) and cleavage of unmethlyated phage DNA thru sequence-specific nucleases (Tock and Dryden, 2005). Finally, suicidal systems, such as abortive infection systems or toxin-antitoxin systems, can be used to abort phage propagation as a sacrifice to protect surrounding clonal population (Dy et al., 2014; Seed, 2015).

Although these defense systems provide innate immune response, a recent discovery of CRISPR-Cas systems showed that prokaryotes also have a sophisticated adaptive immune system. The unique repetitive loci in *E. coli* were first discovered in the late 1980s (Ishino et al., 1987; Nakata et al., 1989) and were named with the acronym CRISPR (clustered regularly interspaced short palindromic repeats) in the early 2000s (Mojica et al., 2000; Jansen et al., 2002). In 2002, a family of CRISPR-associated (*cas*) genes were identified

(Jansen et al., 2002) and in 2005, it was discovered that non repetitive elements within CRISPR loci, or 'spacers', match sequences from phages and plasmids (Bolotin et al., 2005; Mojica et al., 2005; Pourcel et al., 2005). These findings, including the correlation between phage resistance and number of spacers in CRISPR loci (Pourcel et al., 2005), suggested a role for CRISPR-Cas system as an adaptive immune system that functions similar to RNA interference in eukaryotes, using Cas effectors and CRISPR-derived guide RNAs to silence foreign nucleic acids (Makarova et al., 2006).

Four early studies showed the fundamental features of CRISPR-Cas systems as a functional adaptive immune system in prokaryotes. The first study showed that spacers present in *Streptococcus thermophilus* provided resistance against matching phages and that *S. thermophilus* could gain resistance to phages by acquiring new spacers against a newly infecting phage (Barrangou et al., 2007). This study demonstrated that immunity conferred by CRISPR-Cas systems is adaptive. Another study showed that the CRISPR is transcribed and processed to form CRISPR RNAs (crRNAs) that guide a complex of Cas proteins, termed Cascade (CRISPR-associated complex for antiviral defense), which is responsible for immunity against phages in *Escherichia coli* (Brouns et al., 2008). The last two studies discovered that CRISPR-Cas system prevented plasmid conjugation in *Staphylococcus epidermidis* and could target DNA (Marraffini and Sontheimer, 2008), or RNA (Hale et al., 2009). Since then, extensive work has been done to understand the genetics, mechanisms, and applications of CRISPR-Cas system. In this chapter, I will discuss the most recent mechanistic details of CRISPR-Cas systems.

Overview of CRISPR-Cas Systems and Their Diversity

CRISPR-Cas systems are found both in archaea and bacteria. Early estimates suggested that these systems were present in 40 % of bacterial and 90 % of archaeal systems (Sorek et al., 2008; Makarova et al., 2011a); however, later estimates suggested most uncultivated bacteria (~90%) do not contain CRISPR-Cas systems (Burstein et al., 2016). CRISPR loci consist of an AT-rich leader followed by a series of partially palindromic repeat sequences (approximately 30-40 base pairs (bp)) interspaced by short 'spacer' sequences that are mostly derived from phages, plasmids or other mobile genetic elements (Bolotin et al., 2005; Mojica et al., 2005; Pourcel et al., 2005). These loci are usually flanked by accompanying *cas* genes. CRISPR-Cas immunity proceeds in three stages (Fig. 1). The first stage is adaptation, in which a short segment from the viral genome is integrated into the CRISPR locus as a new spacer. During the expression stage, the CRISPR locus is transcribed and processed into short CRISPR RNAs (crRNAs) containing one spacer sequence flanked with partial repeat sequences. The final step is interference, in which the spacer sequence in the crRNA guides Cas effectors for cleavage of the viral genomes bearing complementary sequences to the crRNA spacer sequence (reviewed in Marraffini, 2015; Mohanraju et al., 2016).

CRISPR-Cas systems are hypervariable due to the dynamic co-evolution in the phage-host arms race (Makarova et al., 2015). The systems differ in terms of *cas* gene loci and can be divided into two classes, six types, and many subtypes (Fig. 2) (Makarova et al., 2015; Koonin et al., 2017). Class 1 systems encompass multi-subunit effector complexes composed of Cas proteins in uneven stoichiometry, such as Cascade in type I or Csm/Cmr complexes for type III systems. Although the sequences of protein subunits in type I and type III effector complexes are diverse, the complexes share similarities in their overall

architectures that suggests a common origin (Makarova et al., 2011b; Rouillon et al., 2013; Jackson et al., 2014; Mulepati et al., 2014; Zhao et al., 2014; Jackson and Wiedenheft, 2015; Taylor et al., 2015). However, despite these structural similarities, type I and type III systems are mechanically distinct. Type I effector complexes target double-stranded DNA (Westra et al., 2012; Hochstrasser et al., 2014, 2016; Beloglazova et al., 2015; Elmore et al., 2015; Plagens et al., 2015; Rollins et al., 2015; van Erp et al., 2015; Xiao et al., 2018), while type III complexes target both RNA and transcriptionally active DNA (Hale et al., 2009; Deng et al., 2013; Goldberg et al., 2014; Samai et al., 2015; Jiang et al., 2016b; Kazlauskiene et al., 2016). Also, recent studies showed that, upon binding to target RNA, type III interference complexes function as a cyclic oligoadenylate synthetase that converts ATP into cyclic adenylates to activate other Cas RNases for degradation of nonspecific RNA (Kazlauskiene et al., 2017; Niewoehner et al., 2017).

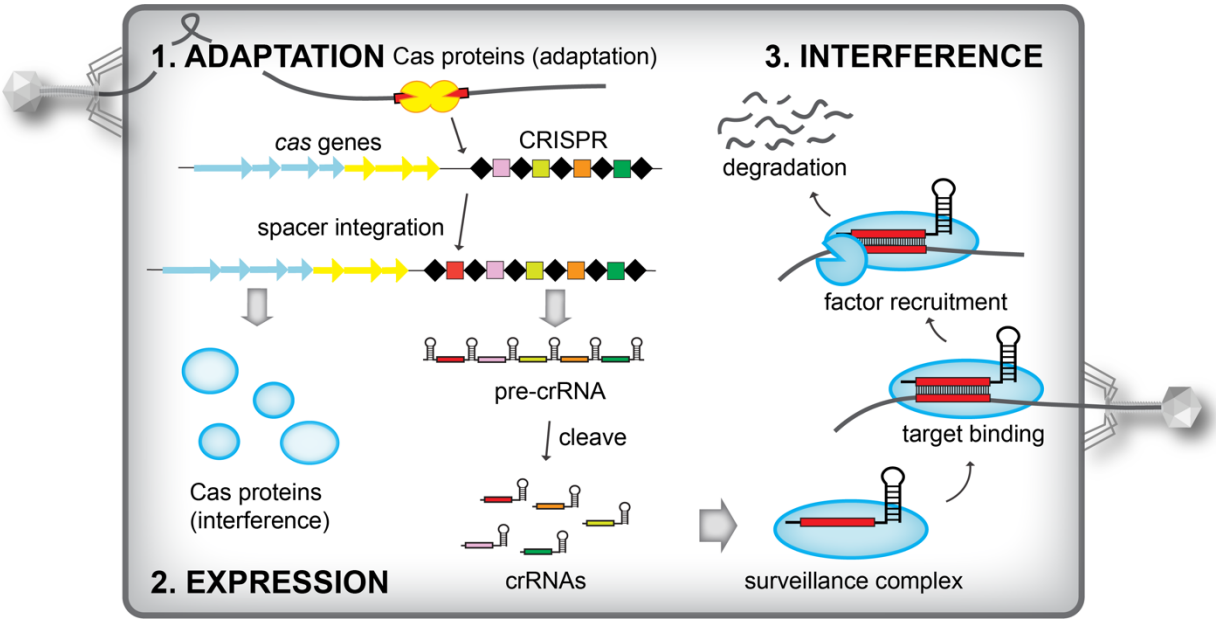

Figure 1. Overview of CRISPR-Cas systems. During the adaptation stage, the adaptation Cas proteins capture and insert short fragments from the viral genomes into the CRISPR locus. In the expression stage, this locus is transcribed and processed into crRNAs. In the interference stage, crRNAs form a surveillance complex with Cas proteins to target the viral genomes that are complementary to the crRNA sequences. Target binding triggers Cas nuclease for degradation.

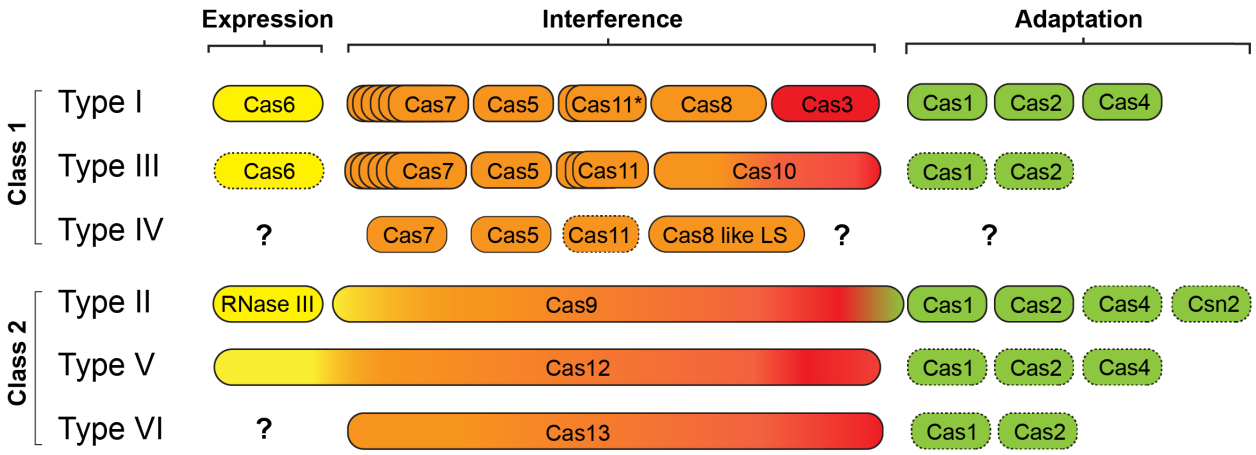

Figure 2. Classification of CRISPR-Cas systems. Class 1 encodes multi-subunit protein complexes while Class 2 uses only a single protein for interference. In several type I subtypes, Cas11 subunits are found as a fusion with Cas8, indicated as an asterisk. Type IV encodes a Cas8-like large subunit. Dashed lines indicate that most systems lack these genes and use Cas proteins provided in trans from other CRISPR-Cas loci. Adapted from Mohanraju et al., 2016.

Class 2 systems encode one single multidomain effector protein, such as Cas9 in type II, Cas12 in type V or Cas13 in type VI systems (Makarova et al., 2015; Koonin et al., 2017). These effector proteins show differences in target recognition, as Cas9 and Cas12 target dsDNA while Cas13 targets RNA (Garneau et al., 2010; Gasiunas et al., 2012; Jinek et al., 2012; Abudayyeh et al., 2016; Shmakov et al., 2016; Liu et al., 2017). Cas9 cleaves each strand of target DNA in a concerted manner using two separate active sites, creating a blunt double-stranded break (DSB). Cas12 uses only a single active site to cut each strand and creates a staggered cut with a 5-nt 5′-overhang (Anders et al., 2014; Jinek et al., 2014; Sternberg et al., 2014, 2015, Jiang et al., 2015, 2016a; Gao et al., 2016; Stella et al., 2017). Because they only require a single protein and guide RNAs for interference, these class 2 proteins have been extensively redesigned for precise genome engineering in many different research fields. By changing the spacers within guide RNA sequences and introducing DSB into a gene of interest, these proteins become a versatile tool for gene manipulation upon repair of the DSB (applications reviewed in Carroll, 2014; Hsu et al., 2014; Barrangou and van Pijkeren, 2016; Zhang et al., 2018).

In contrast to the diversity of interference machinery, most CRISPR-Cas systems contain *cas1* and *cas2* genes that are required for adaptation. The universality of these genes suggests a common molecular mechanism to acquire immunity among all systems. Many systems require additional proteins for spacer acquisition, such as Cas4 in most type I and V systems and type II-B and C, or Csn2 in type II-A systems (Heler et al., 2015; Hudaiberdiev et al., 2017; Koonin et al., 2017). It has been hypothesized that exonucleolytic activity of Cas4 is required for prespacer (e.g. spacer prior to integration) generation (Zhang et al., 2012; Lemak et al., 2013), while Csn2 slides along DNA without any nuclease activity and interacts with Cas9 (Ellinger et al., 2012; Arslan et al., 2013; Ka et al., 2016, 2018). However, the functions of these proteins during spacer acquisition remain elusive.

Type I-C CRISPR-Cas systems in *Bacillus halodurans*

Type I systems are the most widespread and abundant among bacteria and archaea (Makarova et al., 2015) and are defined by the presence of a signature protein Cas3, containing helicase and nuclease domains responsible for degrading the target DNA (Sinkunas et al., 2011; Westra et al., 2012; Mulepati and Bailey, 2013; Hochstrasser et al., 2014; Huo et al., 2014; Xiao et al., 2018). Currently, type I systems can be further divided into 7 subtypes (type I-A through type I-F and I-U) (Koonin et al., 2017). Recent structures showed that type I-C, E, and F systems encode Cascade-like complexes composed of four to seven Cas proteins in uneven stoichiometry, suggesting a similar mechanism in targeting DNA (Jackson et al., 2014; Mulepati et al., 2014; Zhao et al., 2014; Hayes et al., 2016; Hochstrasser et al., 2016; Chowdhury et al., 2017; Guo et al., 2017). Moreover, Cas1 and

Cas2 proteins are conserved among all type I systems, indicating that the spacers are integrated via a universal mechanism.

Our current understanding of type I systems is largely from studies of type I-E systems in *Escherichia coli* K-12 and type I-F systems in *Pseudomonas aeruginosa* and *Pectobacterium atrosepticum*. Type I-C is the second most abundant sub-type among the sequenced genomes of bacteria and archaea (Makarova et al., 2015) and it is unique in that it requires only three proteins to form Cascade-like complexes, instead of five proteins in type I-E *E. coli* Cascade (Hochstrasser et al., 2016). Nevertheless, the molecular mechanisms of CRISPR adaptation and interference in type I-C systems are just beginning to be understood.

This thesis focuses on understanding the mechanisms of CRISPR RNA guided immunity in type I-C system in adaptation and silencing. In particular, I have studied the type I-C system found in the soil bacterium *Bacillus halodurans. B. halodurans* encodes five CRISPR loci along with *cas* operons from two different systems: type III-B and type I-C (Fig. 3). Here, I summarize the overview the three stages (adaptation, expression, and interference) of CRISPR-based adaptive immunity in type I systems and discuss the current understanding of type I-C systems.

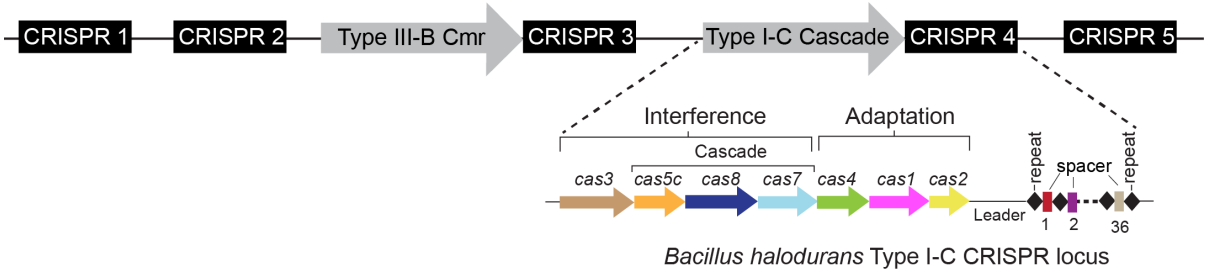

Figure 3. Overview of *cas* genes and CRISPR loci found in *Bacillus halodurans*. Five CRISPR arrays and two different subtypes are found within the genome: type III-B Cmr and type I-C Cascade. Type I-C CRISPR-Cas systems encodes 7 *cas* genes for adaptation and interference. CRISPR locus 4 can be found downstream of I-C subtype specific *cas* genes and contains 36 spacers, which is the largest among the arrays.

Adaptation

CRISPR immunity begins when short segments from foreign nucleic acids are captured and inserted as a molecular memory into a CRISPR locus (reviewed in Sternberg et al., 2016; Jackson et al., 2017). The adaptation stage is fundamental for the subsequent expression and interference stages that neutralize foreign nucleic acids upon re-infection. Although adaptation have been observed in many sub-types (type I-A (Erdmann and Garrett, 2012; Liu et al., 2015), type I-B (Li et al., 2014a, 2014b), type I-C (Rao et al., 2016, 2017), type I-E (Datsenko et al., 2012; Swarts et al., 2012; Yosef et al., 2012), and type I-F (Richter et al., 2014; Vorontsova et al., 2015)), the mechanisms are only partly understood. There are two modes of adaptation: naïve, when the invader has not been previously encountered; and primed, when a pre-existing record from the invader is already present in the CRISPR array (Fig. 4).

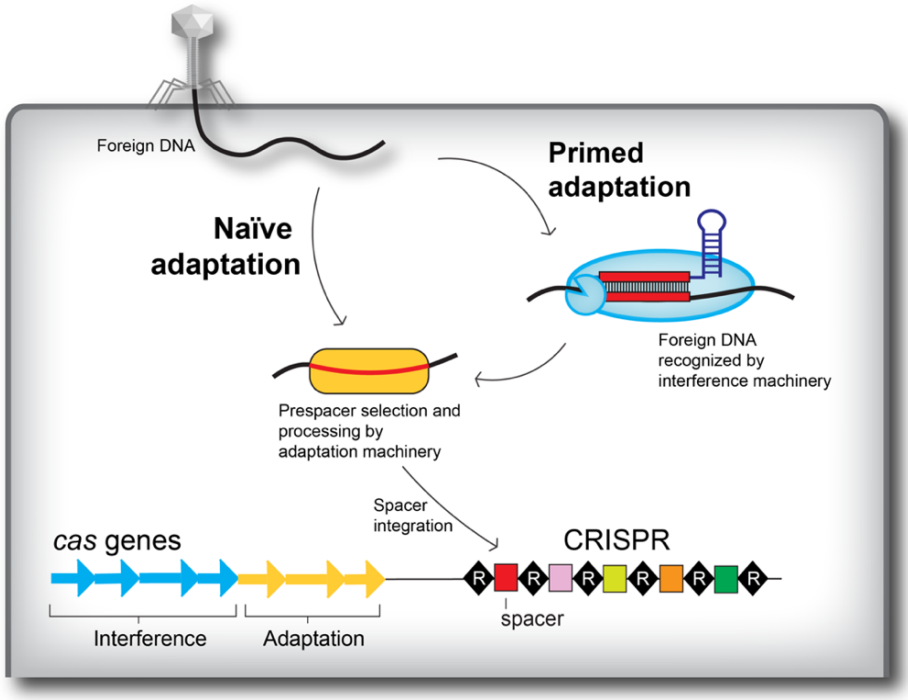

Figure 4. Schematic view of two types of adaptation. Naïve adaptation occurs when there is no information for the target in the CRISPR array and requires only the adaptation machinery. Primed adaptation requires a previous record within the CRISPR array that triggers binding and degradation by the interference machinery and another subsequent spacer integration by the adaptation machinery.

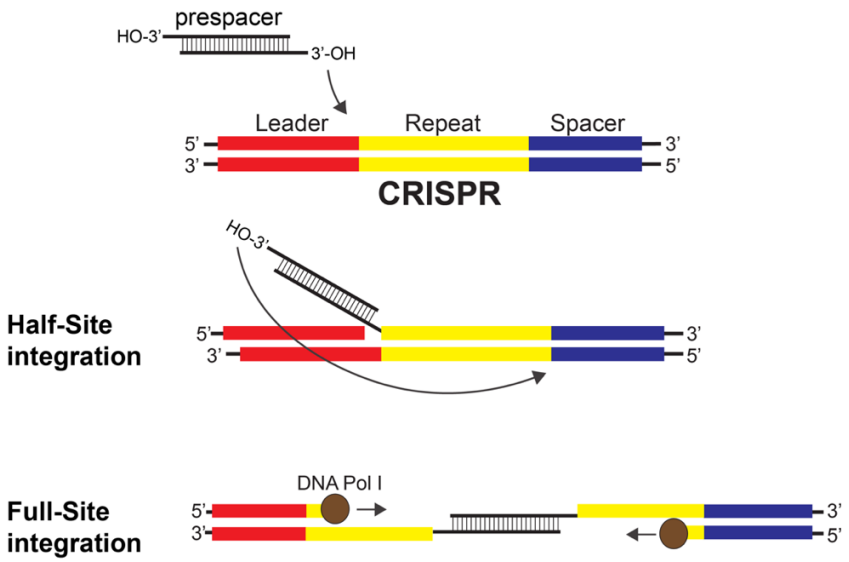

Figure 5. Schematic view of integration step. Adaptation machinery selects and processes prespacers prior to integration. First, the processed prespacer is integrated between the junction of leader and repeat on the positive strand of the CRISPR. Then the other strand of the prespacer attacks the minus strand of the CRISPR at the spacer end of the repeat. Host factors such as a DNA polymerase and a DNA ligase fill the gap and ligate the nick of the gapped-intermediate product.

The key factors for spacer integration are Cas1 and Cas2, which are required for both naive and primed adaptation. An early study of the *E. coli* type I-E system showed that overexpression of Cas1 and Cas2 resulted in newly acquired spacers within the CRISPR array even in the absence of other Cas proteins (Yosef et al., 2012). Later structural studies revealed that *E. coli* Cas1 and Cas2 form a heterohexameric $Cas1_4$-$Cas2_2$ complex (hereafter Cas1-Cas2) that is critical for spacer acquisition (Nuñez et al., 2014). Primed adaptation requires both Cas1-Cas2 and the interference machinery, Cascade and Cas3. Adaptation and interference machineries work together to facilitate rapid spacer acquisition following Cascade-target binding to increase resistance against re-encountered invaders (Datsenko et al., 2012; Swarts et al., 2012; Fineran et al., 2014; Redding et al., 2015; Kunne et al., 2016). In the type I-F system, Cas2 is found to be a fusion with Cas3 forming a Cas1-Cas2/3 complex (Fagerlund et al., 2017; Rollins et al., 2017), suggesting a direct connection between the adaptation and interference machinery. Despite the differences, both naïve and primed

adaptation can be further divided into three steps: the generation of prespacers; the selection and processing of prespacers; and the integration of prespacers into the CRISPR array (Fig. 5).

Generation of prespacers

Given the importance of prespacers for targeting by the surveillance complex during CRISPR interference, the adaptation machinery must precisely select prespacers from invaders. Prespacer substrates from the host DNA must be avoided because this can lead to autoimmunity (Stern et al., 2010; Vercoe et al., 2013). Therefore, to avoid autoimmunity, the prespacer substrates from the invaders must be more abundant than the prespacers from the host DNA. When overexpressing Cas1 and Cas2 in an *E. coli* strain that lacks interference machinery, the acquired spacers are largely from plasmids (e.g. invaders) instead of the host chromosomal DNA (Díez-Villaseñor et al., 2013; Yosef et al., 2013; Nuñez et al., 2014).

Recently, it has been shown that this preference for plasmid DNA is due to a connection between replication forks and spacer acquisition (Levy et al., 2015). During replication, stalled replications forks can create double stranded breaks (DSB), which are repaired in part by the RecBCD complex. The RecBCD complex is composed of helicase and nucleases that unwinds and degrades the DNA back to the nearest Chi site (Dillingham and Kowalczykowski, 2008). It has been suggested that the adaptation machinery uses the degradation products from RecBCD activity for spacer acquisition (Levy et al., 2015) (Fig. 6). Chi sites are overrepresented on *E. coli* chromosome (Colbert et al., 1998) and plasmids replicate more frequently causing more DSBs (Shee et al., 2013), resulting in a greater abundance of RecBCD products from plasmid DNA. Moreover, phages inject linear dsDNA

into the host cell and the linear end is recognized as a DSB and processed by RecBCD (Poranen et al., 2002; Dillingham and Kowalczykowski, 2008). Notably, during adaptation, there is a strong preference for free DNA ends and spacers are acquired immediately during phage DNA injection (Modell et al., 2017; Shiimori et al., 2017). Together, these observations indicate that RecBCD degradation leads to a strong preference for plasmid or phage DNA over the chromosomal DNA during spacer acquisition. However, despite the functional significance of the RecBCD complex during spacer acquisition, strains with deletion of *recB, recC,* or *recD* can still incorporate spacers into the CRISPR array, albeit with a reduced bias towards the foreign DNA (Levy et al., 2015). These results suggest an alternative mechanism for spacer production in the absence of RecBCD.

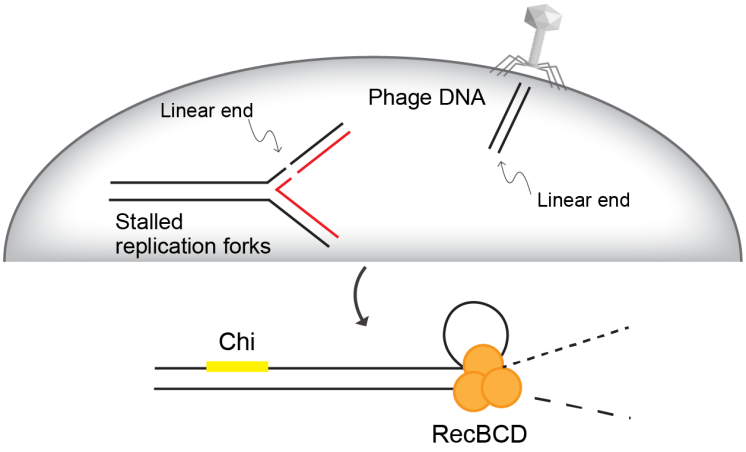

Figure 6. Prespacer substrates production pathway during Naïve adaptation. Stalled replication forks or injected phage DNA as double-strand breaks are processed by RecBCD complex.

Another mechanism for generating prespacers is crRNA-mediated adaptation, a process called priming (Datsenko et al., 2012; Swarts et al., 2012). Priming was first observed in the *E. coli* type I-E system and has since been observed in most other type I systems (Datsenko et al., 2012; Swarts et al., 2012; Savitskaya et al., 2013; Li et al., 2014b; Xue et al., 2015; Staals et al., 2016; Rao et al., 2017). In addition to the adaptation complex, the interference machinery – the crRNA-guided Cascade complex and the nuclease-helicase

Cas3 – are required for priming. Recently, an *in vitro* study showed that the nuclease active Cas3 degrades Cascade-bound target DNA into single-stranded products of 30-100 nucleotides. The fragments were likely to be re-annealed to form partially duplexes, thus facilitating Cas1-Cas2 to integrate the Cas3-derived fragments into the CRISPR locus (Kunne et al., 2016). However, it is unclear whether single stranded fragments from Cas3 and RecBCD activities are re-annealed naturally or with help from other host factors (Fig. 7).

In addition to the degradation products of Cas3, the interference machinery is also thought to promote priming through an interference-independent pathway. Single molecule studies of the type I-E system have shown that, in the presence of Cas1-Cas2, Cas3 nuclease activity is inactivated when it is recruited to Cascade (Redding et al., 2015). Following recruitment, Cas1-Cas2 and Cas3 translocate along the target DNA while remaining bound to Cascade by looping out the DNA, presumably in search of prespacers (Redding et al., 2015; Brown et al., 2017). In type I-F systems, Cas2 is fused with Cas3 and forms a Cas1-Cas2/3 complex (Fagerlund et al., 2017; Rollins et al., 2017). Within this complex, Cas1 inhibits Cas2/3 nuclease activity (Rollins et al., 2017), suggesting a similar mechanism that attenuates interference in favor of nuclease-free translocation when both interference and adaptation machinery are recruited to Cascade. Unlike in type I-E, where Cas1-Cas2-Cas3 translocates unidirectionally and only selects spacers from the target strand, the type I-F Cas1-Cas2/3 complex likely translocates bidirectionally based on the observation of spacers derived from both strands (Savitskaya et al., 2013; Staals et al., 2016). However, the mechanism of how prespacers are excised during this priming process is yet to be resolved.

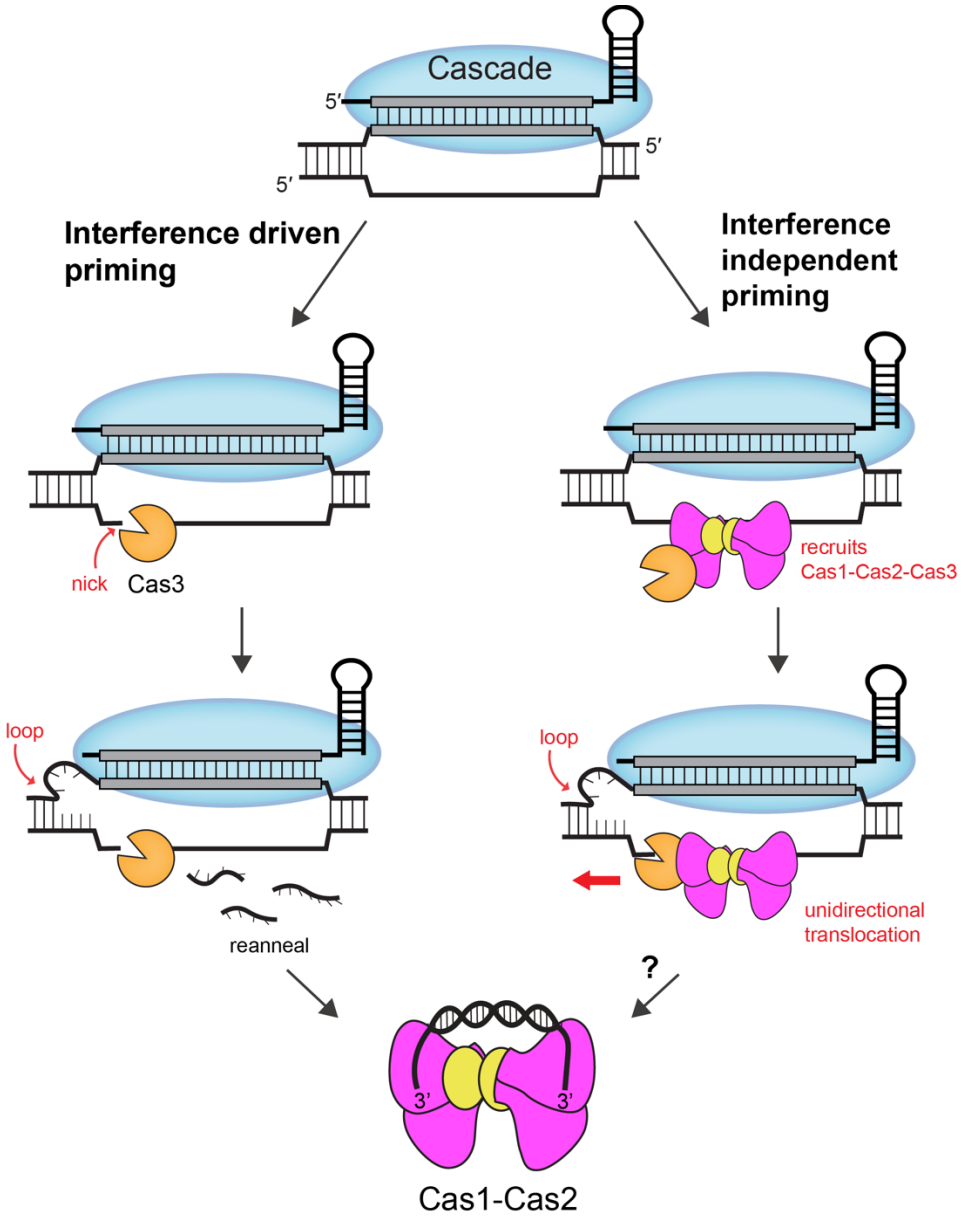

Figure 7. Overview of primed adaptation in *E. coli* type I-E systems. Primed adaptation begins when crRNA-guided Cascade binds to the target. For interference driven pathway, Cascade recruits nuclease active Cas3. The ssDNA products of Cas3 are likely to be reannealed and Cas1-Cas2 uses the substrates for integration. For interference independent pathway, Cascade recruits Cas1-Cas2 and nuclease inactive Cas3. Then Cas1-Cas2-Cas3 translocates unidirectionally along the target to select the prespacers, however it is currently unknown how the prespacers are excised in this process.

Prespacer selection and processing

After the generation of prespacers, the adaptation machinery needs to process the substrate prior to integration. Because many CRISPR-Cas systems have consistent spacer

lengths, it was hypothesized that Cas1-Cas2 must use a ruler-like mechanism (Erdmann and Garrett, 2012; Díez-Villaseñor et al., 2013). Indeed, recent crystal structures revealed that *E. coli* Cas1-Cas2 complexes bound to prespacer dictates the length prior to integration (Nuñez et al., 2015a; Wang et al., 2015) (Fig. 8A). The optimized prespacer contains a 23-bp duplex with 5 nucleotides flanking on both 3′ ends. The duplex is bound along the length of the Cas2 dimer that is sandwiched between two Cas1 dimers. A conserved tyrosine residue on Cas1 subunit is responsible for capping each end of the duplex, while the 3′ single stranded ends of the prespacer are threaded into the active sites of a Cas1 subunit within each dimer. Similar structural constraints have been observed in type I-F Cas1-Cas2/3 and type II Cas1-Cas2 (Fagerlund et al., 2017; Rollins et al., 2017; Xiao et al., 2017a), indicating that Cas1-Cas2 acts like a molecular ruler that predetermines the spacer length prior to integration in many CRISPR-Cas systems.

A critical step in prespacer selection and processing is the recognition of a PAM sequence in the prespacer. The PAM is an important motif that is adjacent to the protospacer and required for interference (Deveau et al., 2008; Semenova et al., 2011). Given the importance of PAM sequences during interference, the adaptation machinery must select prespacers with a canonical PAM and process it at the correct site before integration in order to form functional spacers. Spacers are mostly acquired from a region with a canonical PAM (Savitskaya et al., 2013; Yosef et al., 2013; Staals et al., 2016), indicating that PAMs are recognized by the adaptation machinery during spacer acquisition. For the *E. coli* Cas1-Cas2 complex, PAM recognition occurs within the Cas1 active site (Wang et al., 2015). The crystal structure shows that the PAM complementary sequence (5′-CTT-3′) makes several contacts with residues within the Cas1 active site (Wang et al., 2015) (Fig. 8B). The PAM is not part of the spacer and must be (at least partially) removed from the prespacer substrate

prior to integration. *In vitro* cleavage showed that *E. coli* Cas1 can cleave within the PAM region of the 3′ overhangs, suggesting that the Cas1-Cas2 complex is responsible for prespacer processing in this system. Cas1 cleavage results in a product with free 3′-OH groups on each end, leaving the cytosine from the PAM complementary sequence and generating a final length of 33 bp (Wang et al., 2015).

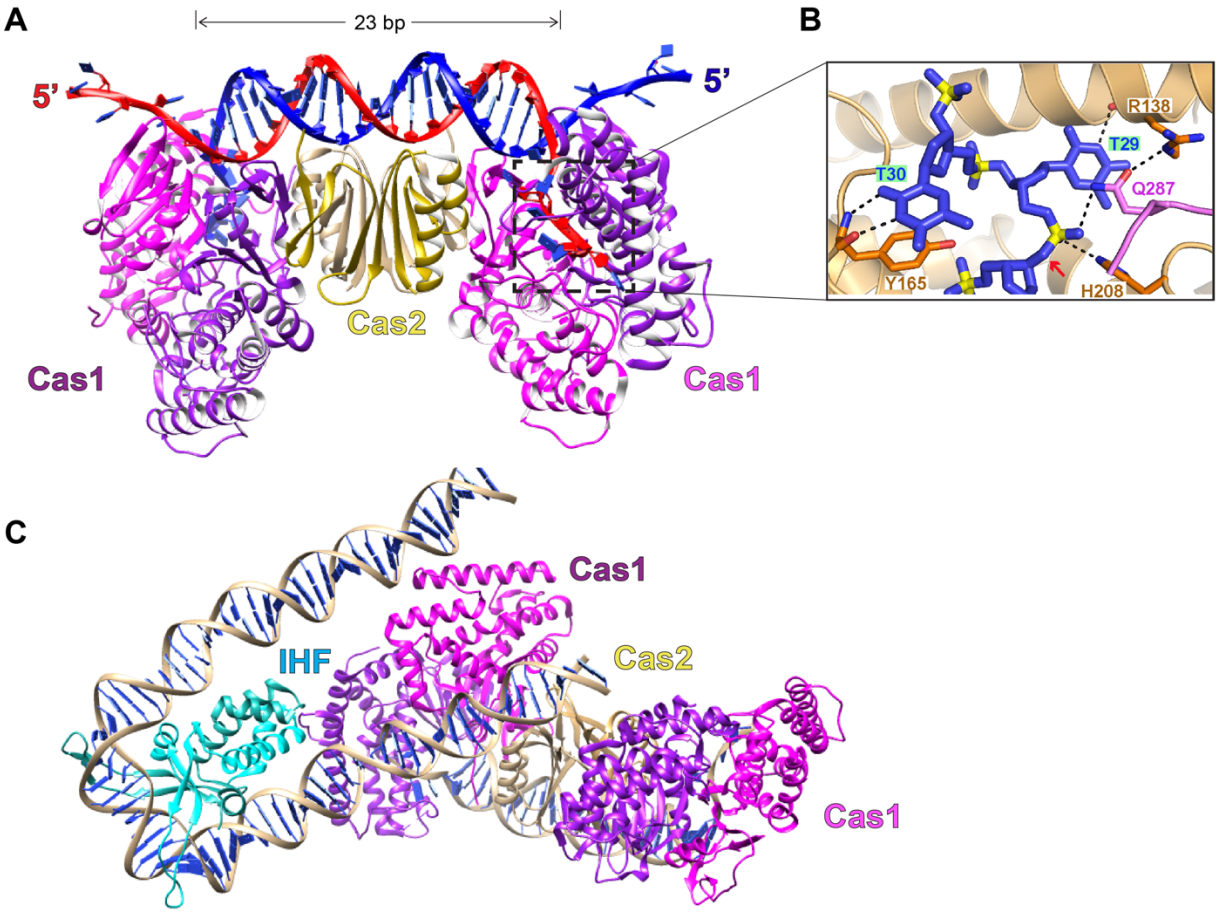

Figure 8. *E. coli* type I-E Cas1-Cas2 complex structures bound to prespacer or with IHF. (A) 2 copies of Cas1 dimer and 2 copies of Cas2 forms a heterohexameric complex with prespacer that are 23 bp duplex with 5 nt on 3′ ends. PDB: 5DQZ (B) The detailed view of sequence specific interactions with Cas1 active sites and PAM complementary sequences. The red arrow indicates the cleavage site. Adapted from Wang., 2015. (C) Structure with Cas1-Cas2 with IHF bound to the extended leader region. PDB: 5WFE

Integration into CRISPR array

After capturing and processing prespacers, Cas1-Cas2 must integrate them precisely within the CRISPR array (Fig. 5). Cas1-Cas2 complex acts as an integrase that resembles retroviral integration and DNA transposition (Nuñez et al., 2015b). Prior to this integration activity, the complex must recognize the CRISPR array to ensure that spacers are not inserted at random sites in the genome. An AT-rich leader is found directly upstream of the CRISPR array and typically spans around 100-500 bp in length (Jansen et al., 2002). The transcriptional promoters for the CRISPR array are located within the leader (Plagens et al., 2012; Carte et al., 2014).

An early *in vivo* study of CRISPR adaptation showed that the leader and a single repeat are sufficient to spacer acquisition (Yosef et al., 2012). Consistently, spacers are specifically acquired at the leader-proximal repeat (Arslan et al., 2014; Nuñez et al., 2015b; Rollie et al., 2015). This polarization provides the chronology of the inserted spacers, where the newest is closer to the leader while the oldest are at the distal end. RNA sequencing data show that the most abundant crRNA species are generated from the spacers from the leader-proximal region of the CRISPR array (Hale et al., 2012). Erroneous integration in the middle of the CRISPR array results in selective pressure against the cells due to low immunity, while leader end integration provides the highest levels of protection to the host (McGinn and Marraffini, 2016). Therefore, the leader specifies the site of integration at the first position of the CRISPR array to provide a more robust immune response against recent invaders.

Several studies have shown that polarized spacer acquisition is governed by intrinsic sequence specificities of Cas1-Cas2 to the leader-repeat region (Rollie et al., 2015; Wright and Doudna, 2016; Xiao et al., 2017a). In some systems, additional factors are required to increase the specificity. For *E. coli* Cas1-Cas2, nonspecific integration was observed *in vitro*

(Nuñez et al., 2015b) and integration host factor (IHF) is required to increase specificity of integration at the leader-repeat junction (Nuñez et al., 2016; Yoganand et al., 2016; Wright et al., 2017). *In vitro* studies showed that IHF induces bending of the leader by ~120º (Yoganand et al., 2016) and this bending promotes integration at the leader-proximal end by 14-fold higher in comparison to the repeat-spacer border (Nuñez et al., 2016). Recent structures of IHF with Cas1-Cas2 bound to a partially integrated prespacer at the leader-repeat junction revealed that IHF-induced leader bending brings Cas1-Cas2 into closer proximity to the upstream of leader region (Wright et al., 2017) (Fig. 8C). While Cas1-Cas2 lacks specific contacts with the leader sequences or IHF, nonspecific interactions induced by IHF are critical to increase the efficiency of integration (Fig. 9).

The recognition of the repeat region by Cas1-Cas2 is also important to integrate spacers properly at the leader-repeat junction. Although repeat types showed weak consistency in both sequence and structure-based classification (Makarova et al., 2015), the repeat usually contains heptameric palindromic sequences or two inverted repeats (IR) that are interspersed with mostly degenerate sequences (Mojica et al., 2000). Mutations of these IR sites inhibits integration activity while mutations of the regions between the two IR sites reduced the activity or impaired maintenance of a constant repeat size (Goren et al., 2016; Wang et al., 2016). However, the structures of type I-E *E. coli* Cas1-Cas2 and type II *E. faecalis* Cas1-Cas2 bound to integration products showed no sequence-specific contacts with the IR sites (Wright et al., 2017; Xiao et al., 2017a). It remains unclear how the repeat is read by Cas1-Cas2 during the dynamic process of integration.

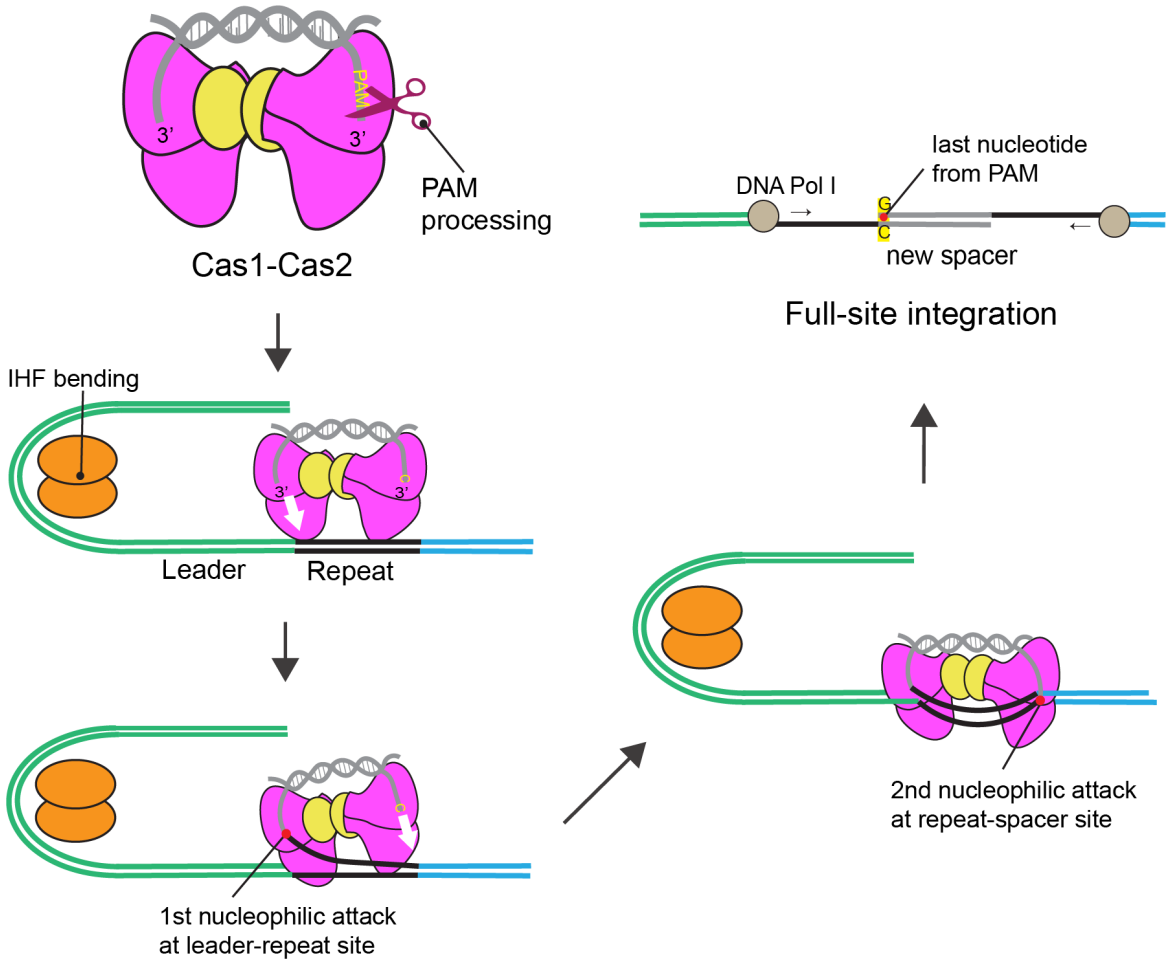

Figure 9. A proposed model of *E. coli* type I-E Cas1-Cas2 integration steps. Cas1-Cas2 processes precisely at PAM sites prior to integration. IHF induces the bending of the leader and Cas1-Cas2 makes a contact in the upstream leader region. Cas1-Cas2 integrates the processed prespacers via two nucleophilic attacks; the first at leader-repeat site and the second at repeat-spacer site. In order to do the second nucleophilic attack, Cas1-Cas2 deforms the repeat region. The newly acquired spacers are flanking with guanosine at 5′ ends. After full site integration, ssDNA gaps are filled and ligated by the host enzymes, DNA polymerase I and DNA ligase.

After initial identification of the leader-repeat junction by Cas1-Cas2, integration occurs thru a two-step integration mechanism. Cas1-Cas2 catalyzes two transesterification reactions through the nucleophilic attack of the 3′-OH groups on each end of the prespacer at the phosphodiester backbone on opposite strands and opposite ends of the first repeat (Arslan et al., 2014; Nuñez et al., 2015b) (Fig. 9). The two attacks occur at the leader-repeat junction on the plus strand and the repeat-spacer junction on the minus strand. Leader-side integration is thought to occur first based on several studies that found that disrupting leader-side

integration also inhibited spacer-side integration (Rollie et al., 2015; Wright and Doudna, 2016; Xiao et al., 2017a). After the first nucleophilic attack, Cas1-Cas2 must define the second site for integration in order to maintain the constant repeat length. IR sites or repeat sequences, which vary between subtypes, act as anchors to specify the second site integration (Wei et al., 2015; Goren et al., 2016; Wang et al., 2016; Wright et al., 2017).

In *E. coli*, the acquired spacers have guanosine as the first nucleotide, which is derived from the last nucleotide of PAM sequences (Swarts et al., 2012; Savitskaya et al., 2013). This directional specificity is important because disruptions can result in destroying PAM and target recognition during interference. Several *in vitro* studies showed that prespacers flanking with G on the 5′ ends were more likely to be integrated (Nuñez et al., 2015b; Rollie et al., 2015). A recent structure revealed that the unfavorable interactions between Cas1 active site harboring cystosine on 3′ ends and leader-repeat site would explain preferences for spacer orientation (Wright et al., 2017); however, the mechanism remains elusive. Furthermore, it is currently unknown how other type I systems, in which the entire PAM is removed from the prespacer, maintain spacer orientation prior to integration.

Integration results in a gapped intermediate, in which the spacer is flanked on either side by a single-stranded repeat. To complete integration, the ssDNA gaps on the repeats can be filled by a DNA polymerase and DNA ligase from the host (Ivančić-Baće et al., 2015) (Fig. 9).

CRISPR RNA biogenesis and Cascade formation

Following spacer insertion, the new molecular memory encodes a CRISPR RNA (crRNA) that can guide the surveillance complex for CRISPR immunity. Prior to

interference, the crRNA expression and maturation stage is required to produce functional guide molecules (Fig. 1). Despite the diversity in CRISPR-Cas systems, most class 1 types share a common molecular principle in the biogenesis stage and it can be divided into three steps: transcription of pre-crRNA, pre-crRNA processing, and formation of crRNA-mediated interference machinery (Fig. 10). First, a long precursor-crRNA (pre-crRNA) is transcribed from a promoter located in the leader. Next, the pre-crRNA is recognized and processed by a metal-independent endoribonuclease from the Cas6 family, which cleave the repeat sequences (Brouns et al., 2008; Carte et al., 2008; Haurwitz et al., 2010; Hatoum-Aslan et al., 2011; Sashital et al., 2011; Wang et al., 2011; Shao et al., 2016). Despite the extreme sequence diversity, the Cas6 family shares a common structural fold that is important in pre-crRNA binding and endonucleolytic cleavage (Hochstrasser and Doudna, 2015). Several crystal structures from type I-E and I-F reveal that the Cas6 enzymes binds the stem-loop region of pre-crRNA in sequence- and structure-specific manner via a positively charged cleft that is formed by two RNA-recognition motifs (RRM) (Haurwitz et al., 2010; Gesner et al., 2011; Sashital et al., 2011) (Fig. 10). Cas6 cleaves at the base of the stem-loop thru a general acid-base mechanism, leaving 5′ hydroxyls and 2′, 3′-cyclic phosphate (Gesner et al., 2011; Jore et al., 2011; Sashital et al., 2011; Haurwitz et al., 2012). Because Cas6 has high affinity for the cleaved products, the enzyme is single-turnover (Sashital et al., 2011; Sternberg et al., 2012). After cleavage, Cas6 from type I-E and I-F remains bound to the stem-loop region as part of the Cascade complex (Jore et al., 2011; Wiedenheft et al., 2011a; Jackson et al., 2014; Mulepati et al., 2014; Zhao et al., 2014; Chowdhury et al., 2017; Guo et al., 2017).

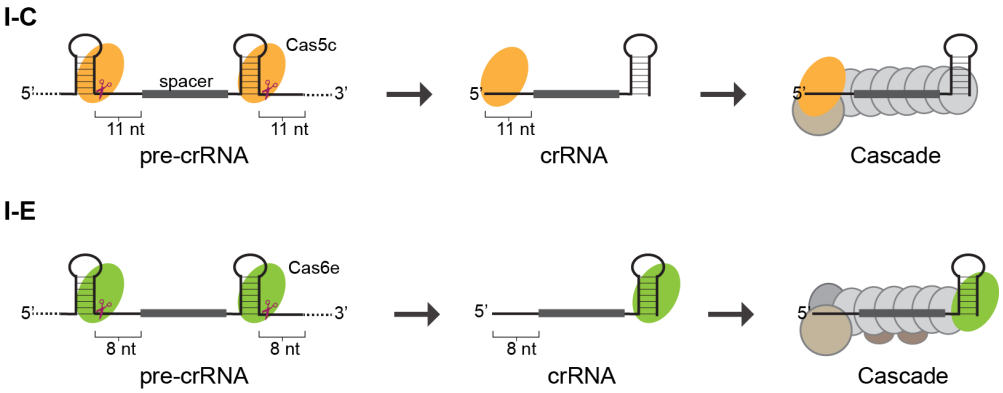

Figure 10. crRNA processing steps in type I-C and type I-E systems. The hairpin structures are recognized by Cas5c in type I-C and Cas6e in type I-E or other type I systems. After cleavage, the matured crRNA retains 11 nt of the repeat for type I-C and 8 nt for type I-E or other type I systems. Cas5c binds to the 5′ ends of crRNA while Cas6e binds to the hairpin on the 3′ end. For complex formation of Cascade, $Cas5c_1$-$Cas7_7$-$Cas8c_1$ are required for type I-C and $Cas5e_1$-$Cas6e_1$-$Cas7_6$-$Cas8e_1$-$Cas11_2$ are required for type I-E.

After processing, mature crRNAs are composed of an 8-nt repeat-derived 5′-handle, invader-derived spacer sequences in the middle, and the stem-loop region of the repeat on the 3′ end (Fig. 10). The crRNAs assemble with Cas proteins in uneven stoichiometry to form Cascade (Jore et al., 2011; Wiedenheft et al., 2011a; Jackson et al., 2014; Mulepati et al., 2014; Zhao et al., 2014). In type I-E Cascade, after cleavage, Cas6e remains bound to the 3′ ends of stem-loop region (Gesner et al., 2011; Sashital et al., 2011). However, Cas6e is dispensable as a component of the Cascade complex when mature crRNAs are provided in a Cas6e-independent manner (Semenova et al., 2015). Next, six copies of Cas7 oligomerize along the crRNA via non-specific interactions. Every 6th position of the crRNA sequence is flipped out due to Cas7 binding and mismatches at these kinked positions do not affect target binding (Jackson et al., 2014; Mulepati et al., 2014; Zhao et al., 2014; van Erp et al., 2015; Hayes et al., 2016). Two copies of Cas11 interact with Cas7 subunits but not directly with the crRNA. Lastly, to complete the formation, Cas5e binds to the 5′ handle of the crRNA and Cas8e contacts Cas5e (Fig. 11A).

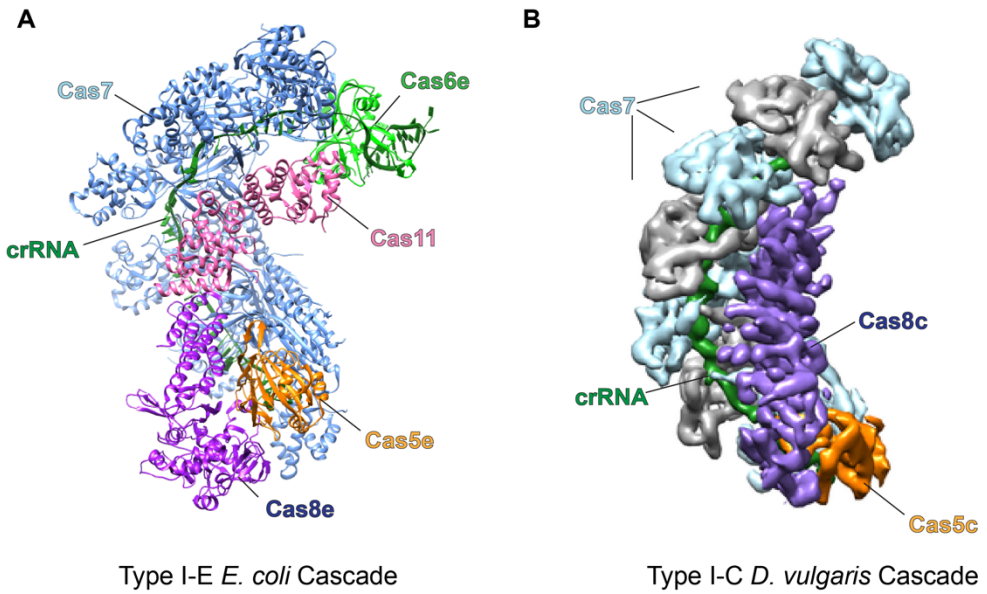

Figure 11. Structures of type I-E *E. coli* Cascade and type I-C *D. vulgaris* Cascade. (A) Crystal structure of Cascade in *E. coli* system. Cascade from type I-E requires $Cas5e_1$-$Cas6e_1$-$Cas7_6$-$Cas8e_1$-$Cas11_2$ and each subunit is indicated and labeled. PDB: 4TVX (B) Cryo EM structure of Cascade in *D. vulgaris* system. Cascade from type I-C requires $Cas5c_1$-$Cas7_7$-$Cas8c_1$. EMDB: 8294

Target Binding and Interference

Once generated, the surveillance complex uses the crRNA as a guide to recognize and trigger degradation of invader nucleic acid. Type I systems target DNA sequences and the nuclease-helicase Cas3 is required for interference (Sinkunas et al., 2011; Westra et al., 2012; Huo et al., 2014; Xiao et al., 2018). The surveillance complex must specifically find its target among the megabases of DNA present in the cell. Cascade simplifies its search by initially recognizing PAM sequences that must be located next to the target (Sashital et al., 2012; Blosser et al., 2015; Redding et al., 2015; Hayes et al., 2016; Xue et al., 2016, 2017). PAM is a short sequence upstream of the target site which is critical to distinguish non-self from self to prevent the system attacking its own genome (Mojica et al., 2009; Semenova et al., 2011; Westra et al., 2013; Rollins et al., 2015). Single molecule studies show that *E. coli* Cascade samples PAM sequences rapidly through three-dimensional diffusion (Redding et

al., 2015; Xue et al., 2017). Cascade interacts with dsDNA transiently with short dwell times (~0.1s) in the absence of PAM, whereas the complex dwells longer at sites of higher PAM density (Xue et al., 2017). As seen in type II and V systems, other crRNA-guided surveillance complexes that must locate targets within dsDNA use this PAM-dependent scanning process as a common mechanism to simplify the search process (Sternberg et al., 2014; Singh et al., 2018).

While scanning for targets, the complex must specifically recognize correct PAM sequences. It has been proposed that several motifs within the Cascade complex may be involved in PAM recognition (Sashital et al., 2012; van Erp et al., 2015; Hayes et al., 2016). Recent structures from *T. fusca* I-E Cascade show that the loop region (L1, residues 130-143) interacts with the PAM (Xiao et al., 2017b), suggesting that it may be responsible for recognizing PAM sequences and unwinding the duplex (Sashital et al., 2012). Moreover, based on the apo- and ssDNA bound Cascade structures, β-hairpin regions on Cas8e and lysine rich loops on Cas7 are necessary for dsDNA binding (van Erp et al., 2015; Van Erp et al., 2018). Notably, *E. coli* Cascade is promiscuous in its PAM recognition (Westra et al., 2012; Fineran et al., 2014) and the structure of Cascade bound to an R-loop mimic shows that the interactions are based on the minor groove interactions between the PAM and the backbone of a glycine residue (G160) on the Cas8e subunit, instead of major groove interactions as observed in type II Cas9 systems (Anders et al., 2014; Hayes et al., 2016). Indeed, mutations within those motifs severely hamper dsDNA binding (van Erp et al., 2015; Xue et al., 2017). Collectively, these regions of Cas8e and Cas7 help to position the dsDNA for PAM recognition during target searching (Sashital et al., 2012; van Erp et al., 2015; Hayes et al., 2016; Xue et al., 2017; Van Erp et al., 2018).

When encountering the correct PAM, Cascade stalls at the site and triggers DNA bending, allowing for the complex to interrogate adjacent DNA for complementarity with the crRNA (Westra et al., 2012; Xiao et al., 2017b). Crystal structures from *E. coli* Cascade have revealed that PAM recognition is coupled to initial duplex destabilization by inserting a glutamine wedge from the Cas8e subunit into the adjacent region of the dsDNA, disrupting the first two base pairs of the seed region and initiating strand separation (Hayes et al., 2016). The seed region (1-5 and 7-8 nt) is located at the PAM proximal end of the protospacer (Semenova et al., 2011; Wiedenheft et al., 2011b; Fineran et al., 2014). After identification of the PAM followed by partial melting and base pairing within the seed region, Cascade forms a seed bubble that eventually promotes the full R-loop structure (Szczelkun et al., 2014; Rutkauskas et al., 2015; Xiao et al., 2017b).

Based on the mechanism of PAM recognition and dsDNA unwinding, Cascade forms R-loops from the PAM-proximal to distal end. Magnetic tweezer experiments have shown that R-loops propagate by zipping the crRNA guide and target sequences and that mismatches can affect the stability of R-loops (Szczelkun et al., 2014; Rutkauskas et al., 2015). Consistently, mutations in the target block interference activity, especially when they occur in the seed region (Semenova et al., 2011; Fineran et al., 2014; Xue et al., 2015; Cooper et al., 2018). Upon reaching a mismatch, R-loop formation stalls, which can result in Cascade dissociation the mismatch is closer to the PAM, or continued zipping to bypass the mismatch. Based on single molecule studies, a mismatch can result in a low fidelity binding mode of Cascade and relatively short-lived R-loops (Blosser et al., 2015; Redding et al., 2015).

Once the complete R-loop is formed to the protospacer end, it creates a steric clash with the Cas11 subunits and triggers a conformational change in Cascade (Wiedenheft et al.,

2011a; Mulepati et al., 2014; Hayes et al., 2016). During this conformational change, the Cas11 dimer slides toward Cas8e and this movement rearranges the C-terminal domain (CTD) of Cas8e. The non-target strand (NTS) is stabilized through interactions with the Cas8e CTD and Cas11 dimer (Xiao et al., 2017b, 2018). A bulge in the NTS facilitates initial nicking by Cas3 (Hayes et al., 2016; Xiao et al., 2017b). The conformational change to stabilize the R-loop occurs after the completion of full R-loop structure (Xiao et al., 2017b). Cas3 is recruited following the Cas8e conformational change in Cascade in the full R-loop state (Westra et al., 2012; Hochstrasser et al., 2014; Xiao et al., 2018). This event suggests that Cascade prevents the premature degradation by Cas3, until the entire protospacer sequence has been checked by the complex (Xiao et al., 2017b). Thus, these multiple checkpoints ensure accuracy during CRISPR interference.

Cas3 contains an N-terminal HD nuclease domain that has ssDNA nuclease activity and C-terminal helicase domain that has metal and ATP-dependent 3′-to-5′ unwinding activity (Sinkunas et al., 2011; Huo et al., 2014). After Cas3 is recruited to the R-loop, Cas3 generates initial nicks at the 7^{th}, 9^{th} and 11^{th} nt from the PAM-proximal end on the NTS in the absence of ATP (Mulepati and Bailey, 2013; Xiao et al., 2017b). However, in the presence of ATP, Cas3 switches modes to processive degradation. Recent structures of Cas3 bound Cascade in pre- and post-nick states have shown the dynamic transitions of Cas3 from nicking to processive degradation modes (Xiao et al., 2018). Cas3 binding does not induce any conformational changes in the Cascade R-loop state; however, the conformational change in the Cas8e CTD that occur upon R-loop stabilization enables Cas3 binding to the Cas8e subunit. This interaction positions Cas3 for cleavage of the NTS bulge, generating an initial nick through the activity of the nuclease domain. After nicking, upon ATP hydrolysis, the nicked NTS is loaded into the helicase domain, and further into the nuclease domain for

processive degradation (Loeff et al., 2018; Xiao et al., 2018). Single molecule studies have shown that Cas3 reels the NTS during degradation, looping out the target strand and generating long ssDNA products based on sporadic nicking activity (Redding et al., 2015; Loeff et al., 2018).

In some cases, mutations in the PAM and target can be tolerated by Cascade binding, but lead to attenuated interference and instead facilitate primed adaptation (Fineran et al., 2014; Semenova et al., 2016). In type I-E, target mutations in close proximity to the PAM cause a conformational change in the Cas8e subunit of Cascade, which inhibits direct recruitment of Cas3 and decreases the rate of target interference (Xue et al., 2016; Krivoy et al., 2018). Instead, it is thought that Cas3 is recruited indirectly for Cas1-Cas2 mediated spacer acquisition upon Cas8e rearrangement (Redding et al., 2015; Xue et al., 2016), although the exact mechanism of this recruitment remains unknown.

Type I-C CRISPR-Cas systems

In this chapter, I have discussed the most recent understanding of CRISPR-Cas immunity, focused on well-characterized type I-E and type I-F systems. The work described in this thesis focuses on the type I-C system, which is less understood. Type I-C shares some similarities with both systems but also has major differences. Here, I discuss our current understanding of type I-C systems.

During adaptation, naïve and primed acquisition events have been observed in *Legionella pneumophila* type I-C system (Rao et al., 2016, 2017), suggesting that adaptation occurs through similar mechanisms to type I-E and I-F. However, unlike type I-E and I-F, all other type I systems including type I-C have an additional adaptation protein,

Cas4, whose role in spacer acquisition has remained mysterious. Cas4 has a RecB nuclease domain and is widespread across types I, II and V (Hudaiberdiev et al., 2017; Koonin et al., 2017). Unlike many other Cas proteins, Cas4 can be found as solo-Cas4 located away from a CRISPR array (Hudaiberdiev et al., 2017), leading to speculation of a function outside of CRISPR defense (Hooton and Connerton, 2015). CRISPR-associated Cas4 has exo- or endo nuclease activities, which varies between different orthologs (Zhang et al., 2012; Lemak et al., 2013, 2014). Thus, it has been hypothesized that Cas4 may provide the prespacer substrates as an alternative to the host RecB nuclease, or that it may be involved in prespacer processing.

My work has focused on elucidating the role of Cas4 in type I-C adaptation. In Chapter 2, we show that Cas4 in the presence of Cas1-Cas2 endonucleolytically processes long 3′ overhang prespacers at PAM site (Lee et al., 2018). Consistent with our work, other studies showed *in vivo* (Kieper et al., 2018; Shiimori et al., 2018) and *in vitro* (Rollie et al., 2017) that Cas4 has a significant role in selecting and processing prespacers. Deleting *cas4* genes reduced the adaptation efficiency and led to integration of longer spacers up to 70 bp in length (Shiimori et al., 2018) or non-functional prespacers mostly from the host genome (Kieper et al., 2018). Collectively, we and others have shown that Cas4 is indispensable for CRISPR-Cas systems to be able to select functional prespacers. In Chapter 3, we showed that Cas4 interacts directly with Cas1-Cas2, forming a Cas4-Cas1-Cas2 complex that mediates spacer selection, processing and integration. We show that within this complex, Cas4 is responsible for PAM recognition and for precise processing just upstream of the PAM in single-stranded substrates.

During the expression stage, rather than Cas6 endoribonuclease found in most of type I systems, type I-C instead uses Cas5c for the generation of mature crRNAs and the integral

subunit of the surveillance complex, Cascade (Garside et al., 2012; Nam et al., 2012; Koo et al., 2013; Punetha et al., 2014; Hochstrasser et al., 2016). Cas5 in other systems does not possess any catalytic functions during immunity but only serves as a structural subunit in Cascade complexes. Similar to Cas6, Cas5c also cleaves the pre-crRNA metal-independently at the base of the stem loop via a general acid/base catalysis reaction. Cas5 uses a catalytic triad of a histidine, tyrosine, and lysine, which function as a general base, a general acid, and to stabilize intermediates, respectively (Garside et al., 2012; Nam et al., 2012; Koo et al., 2013; Punetha et al., 2014). However, unlike Cas6e, Cas5c primarily recognizes the 5′ handle (repeat sequence that remains at the 5′ end after cleavage) and mutation of this region reduces processing activity (Nam et al., 2012; Hochstrasser et al., 2016). Furthermore, Cas5c dissociates more readily from the RNA products (Nam et al., 2012) and has a weaker binding affinity for the crRNA than another Cascade protein, Cas7 (Hochstrasser et al., 2016). But when in complex, Cas5c tightly associates with the crRNA along with other Cas proteins (Hochstrasser et al., 2016), suggesting that pre-crRNA processing and the formation of Cascade may be temporally and spatially coupled (Nam et al., 2012). Crystal structures of Cas5c show an additional extended helical region that may be important in processing, which is not observed in Cas5 from other type I systems (Koo et al., 2013; Jackson et al., 2014). However due to transient interactions with the crRNA, it is difficult to crystalize Cas5c with the RNA substrates (Nam et al., 2012), thus it remains elusive how Cas5c associates with the crRNA.

Whereas other type I systems require four to five Cas proteins, type I-C encodes only three proteins to form its surveillance complex: Cas8c, Cas7 and Cas5c (Fig. 10). However, the overall architecture of type I-C Cascade is similar to *E. coli* Cascade (Hochstrasser et al., 2016) (Fig. 11B). Despite transient interactions between Cas5c and crRNA, the assembly

pathway has been proposed as follows. After cleavage, Cas5c specifically interacts with the 5′ handle of the repeat sequences as soon as seven copies of Cas7 oligomerize along the crRNA up to the 3′ stem-loop. Lastly, Cas8c interacts with Cas5c, Cas7 and crRNA to form type I-C specific Cascade (Hochstrasser et al., 2016). Although *cas8* shows the weakest sequence homology among genes found in type I systems, Cas8c is likely involved in PAM recognition based on the similar orientation of Cas8 in Cascade with the DNA bound (Hayes et al., 2016; Hochstrasser et al., 2016). In addition, Cas8c is thought to be a fusion of Cas8 and Cas11 (Makarova et al., 2015). The C-terminal domain of Cas8c showed structural similarities to Cas11 subunits of type I-E Cascade, which may stabilize the R-loop structure through interactions with the non-target strand (Hochstrasser et al., 2016). Aside from a medium resolution structure of the type I-C Cascade-dsDNA complex, DNA binding by this complex has not been studied extensively. In chapter 4, we show that type I-C Cascade has much higher affinity for non-specific DNA than its counterparts from type I-E and I-F, suggesting an alternative mechanism for target searching involving longer-lived DNA interactions or one-dimensional sliding.

Organization of the dissertation

The aim of this dissertation is to understand the mechanisms of CRISPR RNA guided immunity in the type I-C system during adaptation and interference. In Chapter 2 and 3, we characterized the function of the additional adaptation protein, Cas4, and show that it is responsible for processing prespacers at PAM site. We show that Cas4 directly interacts with Cas1 or Cas1-Cas2 and determine low-resolution structures revealing the architecture of type I-C Cas4-Cas1, Cas1-Cas2 and Cas4-Cas1-Cas2 complexes. In addition, we studied the

effects of sequences within the PAM site for processing. In Chapter 4, we developed the expression and purification of type I-C Cascade for targeting DNA. We showed that type I-C Cascade has much higher affinity to non-specific DNA and initiated the development of a single-molecule FRET assay to visualize how Cascade searches target DNA, which we predict may be different from type I-E Cascade. Lastly, in Chapter 5, I summarized the work and future directions.

References

Abudayyeh, O.O. et al. (2016). C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. Science *5573*.

Anders, C. et al. (2014). Structural basis of PAM-dependent target DNA recognition by the Cas9 endonuclease. Nature.

Arslan, Z. et al. (2013). Double-strand DNA end-binding and sliding of the toroidal CRISPR-associated protein Csn2. Nucleic Acids Res. *41*, 6347–6359.

Arslan, Z. et al. (2014). Detection and characterization of spacer integration intermediates in type I-E CRISPR-Cas system. Nucleic Acids Res. *42*, 7884–7893.

Barrangou, R. et al. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. Science (80-.). *315*, 1709–1712.

Barrangou, R., and van Pijkeren, J.P. (2016). Exploiting CRISPR-Cas immune systems for genome editing in bacteria. Curr. Opin. Biotechnol. *37*, 61–68.

Beloglazova, N. et al. (2015). CRISPR RNA binding and DNA target recognition by purified Cascade complexes from Escherichia coli. Nucleic Acids Res. *43*, 530–543.

Blosser, T.R. et al. (2015). Two distinct DNA binding modes guide dual roles of a CRISPR-cas protein complex. Mol. Cell *58*, 60–70.

Bolotin, A. et al. (2005). Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. Microbiology *151*, 2551–2561.

Breitbart, M., and Rohwer, F. (2005). Here a virus, there a virus, everywhere the same virus? Trends Microbiol. *13*, 278–284.

Brouns, S.J.J. et al. (2008). Small CRISPR RNAs Guide Antiviral Defense in Prokaryotes Stan. Science (80-.). *321*, 960–964.

Brown, M.W. et al. (2017). Assembly and translocation of a CRISPR-Cas primed acquisition complex. Cell 1–13.

Burstein, D. et al. (2016). Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems. Nat. Commun. *7*, 1–8.

Carroll, D. (2014). Genome Engineering with Targetable Nucleases. Annu. Rev. Biochem. *83*, 409–439.

Carte, J. et al. (2008). Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. Genes Dev. *22*, 3489–3496.

Carte, J. et al. (2014). The three major types of CRISPR-Cas systems function independently in CRISPR RNA biogenesis in Streptococcus thermophilus. Mol. Microbiol. *93*, 98–112.

Chibani-chennoufi, S. et al. (2004). Phage-Host Interaction : an Ecological Perspective MINIREVIEW Phage-Host Interaction : an Ecological Perspective. J. Bacteriol. *186*, 3677–3686.

Chowdhury, S. et al. (2017). Structure Reveals Mechanisms of Viral Suppressors that Intercept a CRISPR RNA-Guided Surveillance Complex. Cell *169*, 47–57.e11.

Colbert, T. et al. (1998). Genomics, Chi sites and codons: “Islands of preferred DNA pairing” are oceans of ORFs. Trends Genet. *14*, 485–488.

Cooper, L.A. et al. (2018). Determining the specificity of Cascade Binding, interference, and Primed Adaptation In Vivo in the Escherichia coli Type I-E CRISPR-Cas system. MBio *9*, 1–18.

Datsenko, K.A. et al. (2012). Molecular memory of prior infections activates the CRISPR / Cas adaptive bacterial immunity system. Nat. Commun. *3*, 945–947.

Deltcheva, E. et al. (2011). CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. Nature *471*, 602–607.

Deng, L. et al. (2013). A novel interference mechanism by a type IIIB CRISPR-Cmr module in Sulfolobus. Mol. Microbiol. *87*, 1088–1099.

Deveau, H. et al. (2008). Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. J. Bacteriol. *190*, 1390–1400.

Díez-Villaseñor, C. et al. (2013). CRISPR-spacer integration reporter plasmids reveal distinct genuine acquisition specificities among CRISPR-Cas I-E variants of Escherichia coli. RNA Biol. *10*, 792–802.

Dillingham, M.S., and Kowalczykowski, S.C. (2008). RecBCD Enzyme and the Repair of Double-Stranded DNA Breaks. Microbiol. Mol. Biol. Rev. *72*, 642–671.

Dy, R.L. et al. (2014). Remarkable Mechanisms in Microbes to Resist Phage Infections. Annu. Rev. Virol. *1*, 307–331.

Ellinger, P. et al. (2012). The crystal structure of the CRISPR-associated protein Csn2 from Streptococcus agalactiae. J. Struct. Biol. *178*, 350–362.

Elmore, J. et al. (2015). DNA targeting by the type I-G and type I-A CRISPR–Cas systems of Pyrococcus furiosus. Nucleic Acids Res. gkv1140.

Erdmann, S., and Garrett, R.A. (2012). Selective and hyperactive uptake of foreign DNA by adaptive immune systems of an archaeon via two distinct mechanisms. Mol. Microbiol. *85*, 1044–1056.

Van Erp, P.B.G. et al. (2018). Conformational Dynamics of DNA Binding and Cas3 Recruitment by the CRISPR RNA-Guided Cascade Complex. ACS Chem. Biol. *13*, 481–490.

Fagerlund, R.D. et al. (2017). Spacer capture and integration by a type I-F Cas1–Cas2-3 CRISPR adaptation complex. Proc. Natl. Acad. Sci. 201618421.

Fineran, P.C. et al. (2014). Degenerate target sites mediate rapid primed CRISPR adaptation. Proc. Natl. Acad. Sci. U. S. A. *111*, E1629-38.

Forde, A., and Fitzgerald, G.F. (1999). Bacteriophage defence systems in lactic acid bacteria. Antonie van Leeuwenhoek, Int. J. Gen. Mol. Microbiol. *76*, 89–113.

Gao, P. et al. (2016). Type v CRISPR-Cas Cpf1 endonuclease employs a unique mechanism for crRNA-mediated target DNA recognition. Cell Res. *26*, 901–913.

Garneau, J.E. et al. (2010). The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature *468*, 67–71.

Garside, E.L. et al. (2012). Cas5d processes pre-crRNA and is a member of a larger family of CRISPR RNA endonucleases. RNA *18*, 2020–2028.

Gasiunas, G. et al. (2012). Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. Proc. Natl. Acad. Sci. *109*, E2579–E2586.

Gesner, E.M. et al. (2011). Recognition and maturation of effector RNAs in a CRISPR interference pathway. Nat. Struct. Mol. Biol. *18*, 688–692.

Goldberg, G.W. et al. (2014). Conditional tolerance of temperate phages via transcription-dependent CRISPR-Cas targeting. Nature *514*, 633–637.

Goren, M.G. et al. (2016). Repeat Size Determination by Two Molecular Rulers in the Type I-E CRISPR Array. Cell Rep. *16*, 2811–2818.

Guo, T.W. et al. (2017). Cryo-EM Structures Reveal Mechanism and Inhibition of DNA Targeting by a CRISPR-Cas Surveillance Complex. Cell *171*, 414–426.e12.

Hale, C.R. et al. (2009). RNA-Guided RNA Cleavage by a CRISPR RNA-Cas Protein Complex. Cell *139*, 945–956.

Hale, C.R. et al. (2012). Essential Features and Rational Design of CRISPR RNAs that Function with the Cas RAMP Module Complex to Cleave RNAs. Mol. Cell *45*, 292–302.

Hatfull, G.F. (2008). Bacteriophage genomics. Curr. Opin. Microbiol. *11*, 447–453.

Hatoum-Aslan, A. et al. (2011). Mature clustered, regularly interspaced, short palindromic repeats RNA (crRNA) length is measured by a ruler mechanism anchored at the precursor processing site. Proc. Natl. Acad. Sci. *108*, 21218–21222.

Hatoum-Aslan, A. et al. (2013). A ruler protein in a complex for antiviral defense determines the length of small interfering CRISPR RNAs. J. Biol. Chem. *288*, 27888–27897.

Haurwitz, R.E. et al. (2010). Sequence- and Structure-Specific RNA Processing by a CRISPR Endonuclease. Science (80-.). *329*, 1355–1358.

Haurwitz, R.E. et al. (2012). Csy4 relies on an unusual catalytic dyad to position and cleave CRISPR RNA. EMBO J. *31*, 2824–2832.

Hayes, R.P. et al. (2016). Structural basis for promiscuous PAM recognition in type I–E Cascade from E. coli. Nature 1–16.

Heler, R. et al. (2015). Cas9 specifies functional viral targets during CRISPR-Cas adaptation. Nature *519*, 1–16.

Hochstrasser, M.L. et al. (2014). CasA mediates Cas3-catalyzed target degradation during CRISPR RNA-guided interference. Proc. Natl. Acad. Sci. *111*, 6618–6623.

Hochstrasser, M.L. et al. (2016). DNA Targeting by a Minimal CRISPR RNA-Guided Cascade. Mol. Cell *63*, 840–851.

Hochstrasser, M.L., and Doudna, J.A. (2014). Cutting it close: CRISPR-associated endoribonuclease structure and function. Trends Biochem. Sci. *40*, 58–66.

Hooton, S.P.T., and Connerton, I.F. (2015). Campylobacter jejuni acquire new host-derived CRISPR spacers when in association with bacteriophages harboring a CRISPR-like Cas4 protein. Front. Microbiol. *5*, 1–9.

Hsu, P.D. et al. (2014). Development and applications of CRISPR-Cas9 for genome engineering. Cell *157*, 1262–1278.

Hudaiberdiev, S. et al. (2017). Phylogenomics of Cas4 family nucleases. BMC Evol. Biol. *17*, 232.

Huo, Y. et al. (2014). Structures of CRISPR Cas3 offer mechanistic insights into Cascade-activated DNA unwinding and degradation. Nat. Struct. Mol. Biol. *21*, 771–777.

Ishino, Y. et al. (1987). Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. J. Bacteriol. *169*, 5429–5433.

Ivančić-Baće, I. et al. (2015). Different genome stability proteins underpin primed and naïve adaptation in E. coli CRISPR-Cas immunity. Nucleic Acids Res. *43*, 10821–10830.

Jackson, R.N. et al. (2014). Crystal structure of the CRISPR RNA-guided surveillance complex from Escherichia coli. Science (80-.). *345*, 1473–1479.

Jackson, R.N., and Wiedenheft, B. (2015). A Conserved Structural Chassis for Mounting Versatile CRISPR RNA-Guided Immune Responses. Mol. Cell *58*, 722–728.

Jackson, S.A. et al. (2017). CRISPR-Cas: Adapting to change. Science (80-.). *356*, eaal5056.

Jansen, R. et al. (2002). Identification of genes that are associated with DNA repeats in prokaryotes. Mol. Microbiol. *43*, 1565–1575.

Jiang, F. et al. (2015). A cas9 guide RNA complex preorganized for target DNA recognition. Science *348*, 1477–1482.

Jiang, F. et al. (2016a). Structures of a CRISPR-Cas9 R-loop complex primed for DNA cleavage. Science (80-.). *351*, 867–871.

Jiang, W. et al. (2016b). Degradation of Phage Transcripts by CRISPR-Associated RNases Enables Type III CRISPR-Cas Immunity. Cell *164*, 710–721.

Jinek, M. et al. (2012). A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity. Science (80-.). *337*, 816–821.

Jinek, M. et al. (2014). Structures of Cas9 endonucleases reveal RNA-mediated conformational activation. Science (80-.). *343*.

Jore, M.M. et al. (2011). Structural basis for CRISPR RNA-guided DNA recognition by Cascade. Nat. Struct. Mol. Biol. *18*, 529–536.

Ka, D. et al. (2016). Crystal Structure of Streptococcus pyogenes Cas1 and Its Interaction with Csn2 in the Type II CRISPR-Cas System. Structure *24*, 70–79.

Ka, D. et al. (2018). Molecular organization of the type II-A CRISPR adaptation module and its interaction with Cas9 via Csn2. Nucleic Acids Res. 28–30.

Kazlauskiene, M. et al. (2016). Spatiotemporal Control of Type III-A CRISPR-Cas Immunity: Coupling DNA Degradation with the Target RNA Recognition. Mol. Cell *62*, 295–306.

Kazlauskiene, M. et al. (2017). A cyclic oligonucleotide signaling pathway in type III CRISPR-Cas systems. Science (80-.). *357*, 605–609.

Kieper, S.N. et al. (2018). Cas4 Facilitates PAM-Compatible Spacer Selection during CRISPR Adaptation. Cell Rep. *22*, 3377–3384.

Koo, Y. et al. (2013). Conservation and variability in the structure and function of the Cas5d endoribonuclease in the CRISPR-mediated microbial immune system. J. Mol. Biol. *425*, 3799–3810.

Koonin, E. V. et al. (2017). Diversity, classification and evolution of CRISPR-Cas systems. Curr. Opin. Microbiol. *37*, 67–78.

Krivoy, A. et al. (2018). Primed CRISPR adaptation in Escherichia coli cells does not depend on conformational changes in the Cascade effector complex detected in Vitro. Nucleic Acids Res. *46*, 4087–4098.

Kunne, T. et al. (2016). Cas3-Derived Target DNA Degradation Fragments Fuel Primed CRISPR Adaptation. Mol. Cell 1–13.

Labrie, S.J. et al. (2010). Bacteriophage resistance mechanisms. Nat. Rev. Microbiol. *8*, 317–327.

Lee, H. et al. (2018). Cas4-Dependent Prespacer Processing Ensures High-Fidelity Programming of CRISPR Arrays. Mol. Cell 1–12.

Leenay, R.T. et al. (2016). Identifying and Visualizing Functional PAM Diversity across CRISPR-Cas Systems. Mol. Cell *62*, 137–147.

Lemak, S. et al. (2013). Toroidal structure and DNA cleavage by the CRISPR-associated [4Fe-4S] cluster containing Cas4 nuclease SSO0001 from sulfolobus solfataricus. J. Am. Chem. Soc. *135*, 17476–17487.

Lemak, S. et al. (2014). The CRISPR-associated Cas4 protein Pcal 0546 from Pyrobaculum calidifontis contains a [2Fe-2S] cluster : crystal structure and nuclease activity. Nucleic Acids Res. *42*, 11144–11155.

Levy, A. et al. (2015). CRISPR adaptation biases explain preference for acquisition of foreign DNA. Nature *520*, 505–510.

Li, M. et al. (2014a). Adaptation of the Haloarcula hispanica CRISPR-Cas system to a purified virus strictly requires a priming process. Nucleic Acids Res. *42*, 2483–2492.

Li, M. et al. (2014b). Haloarcula hispanica CRISPR authenticates PAM of a target sequence to prime discriminative adaptation. Nucleic Acids Res. *42*, 7226–7235.

Liu, L. et al. (2017). The Molecular Architecture for RNA-Guided RNA Cleavage by Cas13a. Cell *170*, 714–726.e10.

Liu, T. et al. (2015). Transcriptional regulator-mediated activation of adaptation genes triggers CRISPR de novo spacer acquisition. Nucleic Acids Res. *43*, 1044–1055.

Loeff, L. et al. (2018). Repetitive DNA Reeling by the Cascade-Cas3 Complex in Nucleotide Unwinding Steps. Mol. Cell *70*, 385–394.e3.

Makarova, K.S. et al. (2006). A putative RNA-interference-based immune system in prokaryotes: computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. Biol. Direct *1*, 7.

Makarova, K.S. et al. (2011a). Evolution and classification of the CRISPR-Cas systems. Nat. Rev. Microbiol. *9*, 467–477.

Makarova, K.S. et al. (2011b). Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems. Biol. Direct *6*, 1–27.

Makarova, K.S. et al. (2015). An updated evolutionary classification of CRISPR–Cas systems. Nat. Rev. Microbiol. 1–15.

Marraffini, L.A. (2015). CRISPR-Cas immunity in prokaryotes. Nature *526*, 55–61.

Marraffini, L.A., and Sontheimer, E.J. (2008). CRISPR Interference Limits Horizontal Gene Transfer in Staphylococci by Targeting DNA. Science (80-.). *322*, 1843–1845.

McGinn, J., and Marraffini, L.A. (2016). CRISPR-Cas Systems Optimize Their Immune Response by Specifying the Site of Spacer Integration. Mol. Cell *64*, 616–623.

Modell, J.W. et al. (2017). CRISPR–Cas systems exploit viral DNA injection to establish and maintain adaptive immunity. Nature *10*.

Mohanraju, P. et al. (2016). Diverse evolutionary roots and mechanistic variations of the CRISPR-Cas systems. Science (80-.). *353*, aad5147.

Mojica, F.J.M. et al. (2000). Biological significance of a family of regularly spaced repeats in the genomes of Archaea, Bacteria and mitochondria. Mol. Microbiol. *36*, 244–246.

Mojica, F.J.M. et al. (2005). Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. J. Mol. Evol. *60*, 174–182.

Mojica, F.J.M. et al. (2009). Short motif sequences determine the targets of the prokaryotic CRISPR defence system. Microbiology *155*, 733–740.

Mulepati, S. et al. (2014). Crystal structure of a CRISPR-RNA guided surveillance compex bound to a ssDNA target. Science (80-.). *345*, 1479–1484.

Mulepati, S., and Bailey, S. (2013). In vitro reconstitution of an Escherichia coli RNA-guided immune system reveals unidirectional, ATP-dependent degradation of DNA Target. J. Biol. Chem. *288*, 22184–22192.

Nakata, A. et al. (1989). Unusual nucleotide arrangement with repeated sequences in the Escherichia coli K-12 Chromosome. J. Bacteriol. *171*, 3553–3556.

Nam, K.H. et al. (2012). Cas5d protein processes Pre-crRNA and assembles into a cascade-like interference complex in subtype I-C/Dvulg crispr-cas system. Structure *20*, 1574–1584.

Niewoehner, O. et al. (2017). Type III CRISPR-Cas systems produce cyclic oligoadenylate second messengers. Nature *548*, 543–548.

Nuñez, J.K. et al. (2014). Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. Nat. Struct. Mol. Biol. *21*, 528–534.

Nuñez, J.K. et al. (2015a). Foreign DNA capture during CRISPR–Cas adaptive immunity. Nature *527*, 535–538.

Nuñez, J.K. et al. (2015b). Integrase-mediated spacer acquisition during CRISPR–Cas adaptive immunity. Nature.

Nuñez, J.K. et al. (2016). CRISPR Immunological Memory Requires a Host Factor for Specificity. Mol. Cell 1–10.

Plagens, A. et al. (2012). Characterization of the CRISPR/Cas subtype I-A system of the hyperthermophilic crenarchaeon Thermoproteus tenax. J. Bacteriol. *194*, 2491–2500.

Plagens, A. et al. (2015). DNA and RNA interference mechanisms by CRISPR-Cas surveillance complexes. FEMS Microbiol. Rev. *39*, 442–463.

Poranen, M.M. et al. (2002). Common Principles in Viral Entry. Annu. Rev. Microbiol. *56*, 521–538.

Pourcel, C. et al. (2005). CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. Microbiology *151*, 653–663.

Punetha, A. et al. (2014). Active site plasticity enables metal-dependent tuning of Cas5d nuclease activity in CRISPR-Cas type I-C system. Nucleic Acids Res. *42*, 3846–3856.

Punjani, A. et al. (2017). CryoSPARC: Algorithms for rapid unsupervised cryo-EM structure determination. Nat. Methods *14*, 290–296.

Rao, C. et al. (2016). Active and Adaptive *Legionella* CRISPR-Cas reveals a recurrent challenge to the pathogen. Cell. Microbiol.

Rao, C. et al. (2017). Priming in a permissive type I-C CRISPR-Cas system reveals distinct dynamics of spacer acquisition and loss. RNA *23*, 1525–1538.

Redding, S. et al. (2015). Surveillance and Processing of Foreign DNA by the Escherichia coli CRISPR-Cas System Article Surveillance and Processing of Foreign DNA by the Escherichia coli CRISPR-Cas System. Cell 1–12.

Richter, C. et al. (2014). Priming in the Type I-F CRISPR-Cas system triggers strand-independent spacer acquisition, bi-directionally from the primed protospacer. Nucleic Acids Res. *42*, 8516–8526.

Rollie, C. et al. (2015). Intrinsic sequence specificity of the Cas1 integrase directs new spacer acquisition. Elife *4*.

Rollie, C. et al. (2017). Prespacer processing and specific integration in a Type I-A CRISPR system. Nucleic Acids Res. 1–14.

Rollins, M.F. et al. (2015). Mechanism of foreign DNA recognition by a CRISPR RNA-guided surveillance complex from Pseudomonas aeruginosa. Nucleic Acids Res. *43*, 24–25.

Rollins, M.F. et al. (2017). Cas1 and the Csy complex are opposing regulators of Cas2/3 nuclease activity. Proc. Natl. Acad. Sci. *1*, 201616395.

Rouillon, C. et al. (2013). Structure of the CRISPR interference complex CSM reveals key similarities with cascade. Mol. Cell *52*, 124–134.

Rutkauskas, M. et al. (2015). Directional R-Loop Formation by the CRISPR-Cas Surveillance Complex Cascade Provides Efficient Off-Target Site Rejection. Cell Rep. *10*, 1534–1543.

Samai, P. et al. (2015). Co-transcriptional DNA and RNA Cleavage during Type III CRISPR-Cas Immunity. Cell *161*, 1164–1174.

Sashital, D.G. et al. (2011). An RNA-induced conformational change required for CRISPR RNA cleavage by the endoribonuclease Cse3. Nat. Struct. Mol. Biol. *18*, 680–687.

Sashital, D.G. et al. (2012). Mechanism of Foreign DNA Selection in a Bacterial Adaptive Immune System. Mol. Cell *46*, 606–615.

Savitskaya, E. et al. (2013). High-throughput analysis of type I-E CRISPR/Cas spacer acquisition in E. coli. RNA Biol. *10*, 716–725.

Scheres, S.H.W. (2012). RELION: Implementation of a Bayesian approach to cryo-EM structure determination. J. Struct. Biol. *180*, 519–530.

Seed, K.D. (2015). Battling Phages: How Bacteria Defend against Viral Attack. PLoS Pathog. *11*, 1–5.

Semenova, E. et al. (2011). Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. Proc. Natl. Acad. Sci. *108*, 10098–10103.

Semenova, E. et al. (2015). The Cas6e ribonuclease is not required for interference and adaptation by the E. coli type I-E CRISPR-Cas system. Nucleic Acids Res. 1–13.

Semenova, E. et al. (2016). Highly efficient primed spacer acquisition from targets destroyed by the Escherichia coli type I-E CRISPR-Cas interfering complex. Proc. Natl. Acad. Sci. *113* (27), 7626-7631

Shao, Y. et al. (2016). A Non-Stem-Loop CRISPR RNA Is Processed by Dual Binding Cas6. Structure 1–8.

Shee, C. et al. (2013). Engineered proteins detect spontaneous DNA breakage in human and bacterial cells. Elife, 1–25.

Shiimori, M. et al. (2017). Role of free DNA ends and protospacer adjacent motifs for CRISPR DNA uptake in Pyrococcus furiosus. Nucleic Acids Res. 1–14.

Shiimori, M. et al. (2018). Cas4 Nucleases Define the PAM, Length, and Orientation of DNA Fragments Integrated at CRISPR Loci. Mol. Cell *70*, 814–824.e6.

Shmakov, S. et al. (2016). Discovery and funcional characterization of diverse Class2 CRISPR-Cas systems. Mol. Cell *60*, 385–397.

Singh, D. et al. (2018). Real-time observation of DNA target interrogation and product release by the RNA-guided endonuclease CRISPR Cpf1 (Cas12a). Proc. Natl. Acad. Sci. *115*, 5444–5449.

Sinkunas, T. et al. (2011). Cas3 is a single-stranded DNA nuclease and ATP-dependent helicase in the CRISPR/Cas immune system. EMBO J. *30*, 1335–1342.

Sorek, R. et al. (2008). Phages are the most abundant forms of life on Earth. Nat. Rev. Microbiol. *6*, 181–186.

Staals, R.H.J. et al. (2016). Interference-driven spacer acquisition is dominant over naive and primed adaptation in a native CRISPR–Cas system. Nat. Commun. *7*, 1–13.

Stella, S. et al. (2017). Structure of the Cpf1 endonuclease R-loop complex after target DNA cleavage. Nature *546*, 559–563.

Stern, A. et al. (2010). Self-targeting by CRIPR: gene regulation or autoimmunity? Trends Genet. *26*, 335–340.

Sternberg, S.H. et al. (2012). Mechanism of substrate selection by a highly specific CRISPR endoribonuclease. 661–672.

Sternberg, S.H. et al. (2014). DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. Nature *507*, 62–67.

Sternberg, S.H. et al. (2015). Conformational control of DNA target cleavage by CRISPR-Cas9. Nature *527*, 110–113.

Sternberg, S.H. et al. (2016). Adaptation in CRISPR-Cas Systems. Mol. Cell 1–12.

Swarts, D.C. et al. (2012). CRISPR interference directs strand specific spacer acquisition. PLoS One *7*, 1–7.

Szczelkun, M.D. et al. (2014). Direct observation of R-loop formation by single RNA-guided Cas9 and Cascade effector complexes. Proc. Natl. Acad. Sci. U. S. A. *111*, 9798–9803.

Taylor, D.W. et al. (2015). Structures of the CRISPR-Cmr complex reveal mode of RNA target positioning. Science (80-.). *348*, 581–586.

Tock, M.R., and Dryden, D.T.F. (2005). The biology of restriction and anti-restriction. Curr. Opin. Microbiol. *8*, 466–472.

van Erp, P.B.G. et al. (2015). Mechanism of CRISPR-RNA guided recognition of DNA targets in *Escherichia coli*. Nucleic Acids Res. gkv793.

Vercoe, R.B. et al. (2013). Cytotoxic Chromosomal Targeting by CRISPR/Cas Systems Can Reshape Bacterial Genomes and Expel or Remodel Pathogenicity Islands. PLoS Genet. *9*.

Vorontsova, D. et al. (2015). Foreign DNA acquisition by the I-F CRISPR – Cas system requires all components of the interference machinery. Nucleic Acids Res. 1–13.

Wang, J. et al. (2015). Structural and Mechanistic Basis of PAM-Dependent Spacer Acquisition in CRISPR-Cas Systems. Cell *163*, 840–853.

Wang, R. et al. (2011). Interaction of the Cas6 Riboendonuclease with CRISPR RNAs: Recognition and Cleavage. Structure *19*, 257–264.

Wang, R. et al. (2016). DNA motifs determining the accuracy of repeat duplication during CRISPR adaptation in Haloarcula hispanica. Nucleic Acids Res. *44*, 4266–4277.

Wei, Y. et al. (2015). Sequences spanning the leader-repeat junction mediate CRISPR adaptation to phage in Streptococcus thermophilus. Nucleic Acids Res. *43*, 1749–1758.

Westra, E.R. et al. (2012). CRISPR Immunity Relies on the Consecutive Binding and Degradation of Negatively Supercoiled Invader DNA by Cascade and Cas3. Mol. Cell *46*, 595–605.

Westra, E.R. et al. (2013). Type I-E CRISPR-Cas Systems Discriminate Target from Non-Target DNA through Base Pairing-Independent PAM Recognition. PLoS Genet. *9*.

Wiedenheft, B. et al. (2011a). Structures of the RNA-guided surveillance complex from a bacterial immune system. Nature *477*, 486–489.

Wiedenheft, B. et al. (2011b). RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. Proc. Natl. Acad. Sci. *108*, 10092–10097.

Wright, A. V. et al. (2017). Structures of the CRISPR genome integration complex. Science *0679*, eaao0679.

Wright, A. V., and Doudna, J.A. (2016). Protecting genome integrity during CRISPR immune adaptation. Nat. Struct. Mol. Biol. *23*, 876–883.

Xiao, Y. et al. (2017a). How type II CRISPR–Cas establish immunity through Cas1–Cas2-mediated spacer integration. Nature 1–23.

Xiao, Y. et al. (2017b). Structure Basis for Directional R-loop Formation and Substrate Handover Mechanisms in Type I CRISPR- Cas System Article Structure Basis for Directional R-loop Formation and Substrate Handover Mechanisms in Type I CRISPR-Cas System. Cell *170*, 48–60.e11.

Xiao, Y. et al. (2018). Structure basis for RNA-guided DNA degradation by Cascade and Cas3. Science *0839*, 1–12.

Xue, C. et al. (2015). CRISPR interference and priming varies with individual spacer sequences. Nucleic Acids Res. *43*, 10831–10847.

Xue, C. et al. (2016). Conformational Control of Cascade Interference and Priming Activities in CRISPR Immunity Short Article Conformational Control of Cascade Interference and Priming Activities in CRISPR Immunity. Mol. Cell 1–9.

Xue, C. et al. (2017). Real-Time Observation of Target Search by the CRISPR Surveillance Complex Cascade. Cell Rep. *21*, 3717–3727.

Yoganand, K.N.R. et al. (2016). Asymmetric positioning of Cas1–2 complex and Integration Host Factor induced DNA bending guide the unidirectional homing of protospacer in CRISPR-Cas type I-E system. Nucleic Acids Res. 1–15.

Yosef, I. et al. (2012). Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. Nucleic Acids Res. *40*, 5569–5576.

Yosef, I. et al. (2013). DNA motifs determining the efficiency of adaptation into the Escherichia coli CRISPR array. Proc. Natl. Acad. Sci. *110*, 14396–14401.

Zhang, C. et al. (2018). Development and application of CRISPR/Cas9 technologies in genomic editing. Hum. Mol. Genet. *27*, 1–15.

Zhang, J. et al. (2012). The CRISPR Associated Protein Cas4 Is a 5′ to 3′ DNA Exonuclease with an Iron-Sulfur Cluster. PLoS One *7*.

Zhao, H. et al. (2014). Crystal structure of the RNA-guided immune surveillance Cascade complex in Escherichia coli. Nature.

CHAPTER 2. CAS4-DEPENDENT PRESPACER PROCESSING ENSURES HIGH-FIDELITY PROGRAMMING OF CRISPR ARRAYS

Hayun Lee, Yi Zhou, David W. Taylor, Dipali G. Sashital

(Published in Molecular Cell (2018) 70 (1): 48-59)

Abstract

CRISPR-Cas immune systems integrate short segments of foreign DNA as spacers into the host CRISPR locus to provide molecular memory of infection. Cas4 proteins are widespread in CRISPR-Cas systems and are thought to participate in spacer acquisition, although their exact function remains unknown. Here we show that *Bacillus halodurans* type I-C Cas4 is required for efficient prespacer processing prior to Cas1-Cas2 mediated integration. Cas4 interacts tightly with the Cas1 integrase, forming a heterohexameric complex containing two Cas1 dimers and two Cas4 subunits. In the presence of Cas1 and Cas2, Cas4 processes double-stranded substrates with long 3′-overhangs through site-specific endonucleolytic cleavage. Cas4 recognizes PAM sequences within the prespacer and prevents integration of unprocessed prespacers, ensuring that only functional spacers will be integrated into the CRISPR array. Our results reveal the critical role of Cas4 in maintaining fidelity during CRISPR adaptation, providing a structural and mechanistic model for prespacer processing and integration.

Introduction

In bacteria and archaea, clustered regularly interspaced short palindromic repeats (CRISPR) and CRISPR associated (Cas) proteins provide an adaptive immune system

against mobile genetic elements (Barrangou et al., 2007; Brouns et al., 2008; Marraffini and Sontheimer, 2008). A CRISPR array consists of a series of direct repeats that are flanked by short sequences derived from a foreign genome, called spacers (Bolotin et al., 2005; Mojica et al., 2005; Pourcel et al., 2005). CRISPR-Cas immunity proceeds through three stages: adaptation, expression and interference (reviewed in Marraffini, 2015; Mohanraju et al., 2016)). During adaptation, small fragments of foreign DNA are captured and integrated as new spacers into the CRISPR array by the Cas1-Cas2 complex (Yosef et al., 2012; Nuñez et al., 2014, 2015a; Wang et al., 2015; Jackson et al., 2017; Xiao et al., 2017). During the expression stage, the array is transcribed and processed into short CRISPR RNAs (crRNAs), which assemble with Cas proteins to form a RNA-guided surveillance complex (Brouns et al., 2008; Carte et al., 2008; Haurwitz et al., 2010; Deltcheva et al., 2011; Gesner et al., 2011; Sashital et al., 2011; Hatoum-Aslan et al., 2013; Hochstrasser and Doudna, 2015; Jackson and Wiedenheft, 2015). Finally, during the interference stage, the surveillance complex recognizes targets, often by locating a protospacer adjacent motif (PAM) that can be found immediately next to target protospacer sequence (Mojica et al., 2009; Semenova et al., 2011; Sternberg et al., 2014; Redding et al., 2015; Xue et al., 2017). Following complementary base pairing between the crRNA and protospacer, a Cas nuclease degrades the target and neutralizes the infection (Barrangou et al., 2007; Brouns et al., 2008; Marraffini and Sontheimer, 2008; Garneau et al., 2010; Westra et al., 2012).

CRISPR systems can be classified into two classes, six types (types I-VI), and many subtypes based on the architecture and composition of their *cas* gene loci (Koonin et al., 2017). Despite this divergence, Cas1 and Cas2 are conserved among all CRISPR systems, suggesting that spacers are acquired via a universal mechanism. Cas1 functions as an integrase while Cas2 provides a structural scaffold that enhances the integration activity of

Cas1 (Yosef et al., 2012; Nuñez et al., 2014, 2015a; Wang et al., 2015; Xiao et al., 2017). The Cas1-Cas2 integrase complex captures prespacers that are flanked by 3′-hydroxyl groups on each end and catalyzes the integration reaction at the leader-proximal repeat through direct nucleophilic attack (Nuñez et al., 2015b). The A-T rich leader is found directly upstream of the repeat-spacer array and provides polarized spacer acquisition that is governed by the intrinsic sequence specificity of Cas1-Cas2 for leader-specific sites (Rollie et al., 2015; McGinn and Marraffini, 2016; Wright and Doudna, 2016; Xiao et al., 2017), and enables rapid defense against the most recent invader. The type I-E Cas1-Cas2 complex additionally relies on association with the integration host factor (IHF), which induces DNA bending within the leader and provides additional sequence specificity to Cas1-Cas2 for the leader-repeat junction (Nuñez et al., 2016; Wright et al., 2017).

Although Cas1 and Cas2 are universally required for spacer integration, other type-specific *cas* genes have also been implicated in adaptation (Koonin et al., 2017). In particular, Cas4 is a core family of Cas proteins present in several sub-types within type I, II and V systems (Hudaiberdiev et al., 2017). *In vivo* studies have shown that deletion of the *cas4* gene prevents the uptake of new spacers (Li et al., 2014). Cas4 contains four conserved cysteine residues that coordinate an iron-sulfur cluster and RecB-like nuclease motifs that are required for DNA binding and cleavage activity (Lemak et al., 2013, 2014). Biochemical studies have revealed that Cas4 exhibits DNA unwinding, exo- and endonuclease activity, although the exact activity varies between different orthologs (Zhang et al., 2012; Lemak et al., 2013, 2014). Based on this nuclease activity, it has been hypothesized that Cas4 is involved in spacer generation for Cas1-Cas2 meditated integration, and recent evidence suggests that Cas4 nuclease activity may trim the ends of precursor integration substrates (Rollie et al., 2017).

Of the core Cas family proteins, Cas4 remains one of the few for which the mechanistic role in CRISPR-Cas immunity remains poorly understood. Here, we show that Cas4 plays an integral role in prespacer processing prior to integration by the Cas1-Cas2 complex. Cas4 processes long 3′-DNA overhangs on precursor substrates through Cas1-Cas2-dependent endonucleolytic activity, generating correctly sized substrates prior to integration. The adaptation complex selectively processes pre-spacers with correct PAM sequences, ensuring that only functional spacers are captured during acquisition. While the Cas1-Cas2 complex integrates longer precursor substrates in the absence of Cas4, the presence of Cas4 prevents premature integration and promotes preferential integration of only processed prespacers into the CRISPR locus. Combined with structural analysis of the Cas4-Cas1 complex, these biochemical results indicate that the Cas4 and Cas1 active sites compete for single-stranded overhangs, and that cleavage by Cas4 is prerequisite to integration within the full adaptation complex. Overall, these findings reveal the role of Cas4 in prespacer generation and in ensuring the fidelity of spacer length and PAM selection during spacer integration.

Results

Complex formation by type I-C Cas4, Cas1 and Cas2

Unlike the well-studied type I-E and I-F systems, other type I systems, including the widespread type I-C, contain *cas4* genes. We wondered whether Cas4 from the type I-C system interacts with either Cas1 or Cas2, or with the Cas1-Cas2 complex. Cas4 is found as a fusion with Cas1 in some systems (Hudaiberdiev et al., 2017), and Cas4 from type I-A has previously been shown to interact with a Cas1/2 fusion and the sub-type specific Csa1

protein (Plagens et al., 2012). However, this reconstitution was only achieved upon refolding of denatured proteins, and it is unclear whether the native proteins interact. To test potential interactions under native conditions, we co-expressed Cas1, Cas2, and Cas4 from the *B. halodurans* type I-C system in *E. coli*. Although Cas2 did not co-purify with the complex, an amino-terminal poly-histidine-tagged Cas4 co-purified with untagged Cas1 and the complex was maintained when the affinity tag was moved to the amino terminus of Cas1 (Fig. 1A-B). Regardless of which subunit was tagged, the two proteins formed a stable complex that eluted as a single peak on a size exclusion column with an estimated size of ~150 kDa (Fig. 1C). These results demonstrate that Cas4 directly interacts with Cas1 under native folding conditions, and that the two proteins form a stable complex.

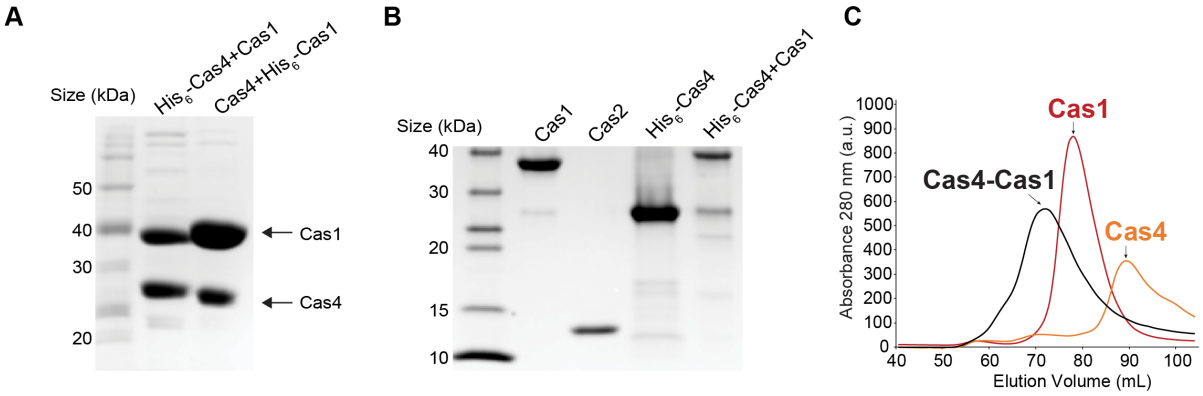

Figure 1. *B. halodurans* Cas1, Cas2, Cas4, and the Cas4-Cas1 complex. (A) Co-expressed and purified Cas4-Cas1 complex on 12% SDS PAGE gel stained with Coomassie blue. His6-Cas4 with untagged Cas1 or His6-Cas1 with untagged Cas4 were co-expressed and complex was purified by nickel affinity chromatograph and size-exclusion chromatography. The stoichiometry of the two preparations are different due to incomplete separation of Cas4-His6-Cas1 complex from free His6- Cas1 on size-exclusion column. Therefore, His6-Cas4-Cas1 was used for all biochemistry and structural studies. (B) Coomassie-blue stained SDS/PAGE gel of purified proteins used in this study. (C) Size- exclusion chromatography (SEC) of co-purified Cas4-Cas1 complex with individually purified Cas1 and Cas4.

To investigate whether Cas1 or the Cas4-Cas1 complex interacts with Cas2, we incubated poly-histidine tagged Cas1 or Cas4-Cas1 complex and untagged Cas2 and performed pull-down assays using nickel affinity chromatography. The untagged Cas2 alone eluted at low imidazole concentrations, whereas some Cas2 co-eluted with Cas1 or Cas4-

Cas1 at high imidazole concentrations (Fig. 2). These data suggest that both Cas1 and Cas4-Cas1 form higher-order complexes with Cas2, supporting a role for Cas4 within the adaptation machinery. However, the complexes do not appear to have the expected stoichiometry (Nuñez et al., 2014), and Cas2 partially eluted at low imidazole concentrations (Fig. 2). Moreover, neither the Cas1-Cas2 nor the Cas4-Cas1-Cas2 complex could be co-purified by size exclusion chromatography. These data suggest that interactions between Cas1 and Cas2 are relatively weak for these orthologs, unlike Cas1 and Cas2 from other sub-types (Nuñez et al., 2015a; Wang et al., 2015; Fagerlund et al., 2017; Rollins et al., 2017; Xiao et al., 2017).

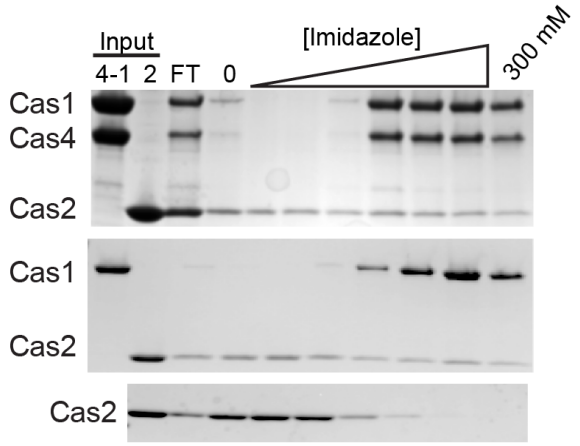

Figure 2. Nickel affinity pull-down of poly-histidine-tagged Cas4-Cas1 complex or His_6-Cas1 and untagged Cas2. A stepwise elution using an imidazole titration (20-300 mM) was performed. FT: Flow through. Untagged Cas2 alone was used as a control (bottom panel).

Molecular architecture of the Cas4-Cas1 complex

To characterize the molecular architecture of the Cas4-Cas1 complex, we performed single-particle electron microscopy (EM) of negatively stained Cas4-Cas1 complexes. Raw micrographs showed globular, monodispersed particles with internal structural features. Two rounds of reference-free two-dimensional (2D) alignment and classification were performed to remove low quality particles, resulting in a total of ~13,000 particles that were used for

further analysis. Distinct 2-fold symmetry was observed in many of the 2D class averages. A 3D model generated by ab initio 3D reconstruction in cryoSPARC (Punjani et al., 2017) was used as an initial model for 3D classification using 3 classes in RELION (Scheres, 2012). 4,590 structurally homogenous particles were extracted from the best 3D model for iterative refinement of the model with imposed C2 symmetry, which led to a final 3D reconstruction of the Cas4-Cas1 complex at ~21 Å resolution using the gold standard 0.143 criterion.

The 2-fold symmetric Cas4-Cas1 complex resembles a crab-claw and is ~130 Å in the longest dimension and ~90 Å wide, with four distinct subunits (Fig. 3A). Notably, the crystal structure of the Cas4 protein Pcal_0546 from *Pyrobaculum calidifontis* (PDB ID: 4R5Q) (Lemak et al., 2014) fits perfectly into the two small subunits at the top of the claw, while the crystal structure of the Cas1 dimer in the Cas1-Cas2 complex from *E. coli* (PDB ID: 4P6I) (Nuñez et al., 2014) were easily accommodated into the two larger subunits at the base of the claw in the final 3D reconstruction (Fig. 3B), suggesting a stoichiometry of 4 Cas1 and 2 Cas4 proteins for the Cas4-Cas1 complex. Docking of these crystal structures into our map places the Cas4 monomer directly above the Cas1 monomer, with the active site residues of K138 and H208 for Cas4 and Cas1, respectively, roughly parallel to each other within the complex (Fig. 3C).

Interestingly, although possessing similar stoichiometry to the Cas4-Cas1 complex, the Cas1-Cas2 complex has a remarkably different molecular architecture (Nuñez et al., 2014). The distances between Cas1 dimers is smaller in the Cas4-Cas1 complex as compared to the Cas1 dimers in *E. coli* Cas1-Cas2 complex with distances between active site residues H208 of 67 and 84- Å for each complex, respectively (Fig. 3D-E). Thus, it is unlikely that two Cas2 molecules could be accommodated in the interface between Cas1 dimers without significantly altering the conformation of Cas4-Cas1. Notably, EM micrographs of Cas4-

Cas1-Cas2 complex co-eluted from nickel-affinity purification (Fig. 2) showed particles of a larger size than for Cas4-Cas1. However, attempts at 3D reconstruction from these images proved unsuccessful, potentially because of sample heterogeneity due to incomplete formation of the putative Cas4-Cas1-Cas2 complex.

Cas4 enhances prespacer processing

The previously demonstrated nuclease activities of Cas4 suggest that it may be involved in prespacer processing prior to integration. The design of the short duplex substrates was based on previous crystal structures of *E. coli* Cas1-Cas2 complex bound to prespacers and the average length (34.4 bp) of the 35 spacers found in the *B. halodurans* CRISPR locus 4 (Nuñez et al., 2015a; Wang et al., 2015). We investigated the effect of temperature on processing activity by incubating reactions either at 37 or 65°C. *Bacillus* strains are facultive alkaliphiles and specifically *B. halodurans* strains are polyextremophiles to temperature up to 90-100 °C and high salt concentrations (Smaali et al., 2006; Dua and Gupta, 2017). Notably, while we observed cleavage of the 3′-overhang substrates at 65ºC, we did not observe cleavage of 5′-overhang or blunt end substrates under similar conditions. Exonucleolytic cleavage of the blunt end substrate was observed using free Cas4, but only at very high concentrations. We therefore proceeded with experiments testing cleavage of substrates with long 3′ overhangs (Fig. 4A).

Incubation of Cas1, Cas2 and an unprocessed prespacer bearing 15-nt 3′ overhangs generated a small amount of cleaved products consistent with the length of a processed prespacer with short overhangs on the 3′ ends (Fig. 4B, lane 6). Strikingly, the processing activity was enhanced substantially in the presence of Cas4 (Fig. 4B, lanes 7-8, Fig. 4D), suggesting that Cas4 may be directly involved in prespacer cleavage. No cleaved products

were observed when Cas1, Cas2, Cas4 or the Cas4-Cas1 complex were incubated individually with DNA (Fig. 4B, lanes 2-5), indicating that optimal processing activity requires all three adaptation proteins. In addition, overhang and duplex length had little effect on overall processing activity both in the absence and presence of Cas4.

To determine the extent to which each subunit of the Cas4-Cas1 complex contributes to the catalytic activity of prespacer processing, we introduced mutations in the active sites of each subunit (Fig. 4C). Based on sequence alignments, Cas4 Lys110 is located in one of the conserved RecB nuclease motifs (motif III) and Cas1 His234 is found in the predicted active site as reported in *E. coli* and *S. pyogenes* (Nuñez et al., 2015b; Wright and Doudna, 2016). While the Cas1 active site mutant (H234A) ablated Cas1-Cas2 processing activity (Fig. 4C, lane 10), addition of individually purified Cas4 or Cas4-Cas1 complex containing H234A Cas1 restored cleavage to wild-type levels (Fig. 4C, lanes 11-12, Fig. 4D). In contrast, the Cas4-Cas1 complex containing K110A Cas4 showed no detectable products (Fig. 4C, lane 13, Fig. 4D). Together these data reveal that while the Cas1 active site can catalyze low levels of prespacer processing in the absence of Cas4, the Cas4 catalytic site is both necessary and sufficient for processing when all three adaptation proteins are present.

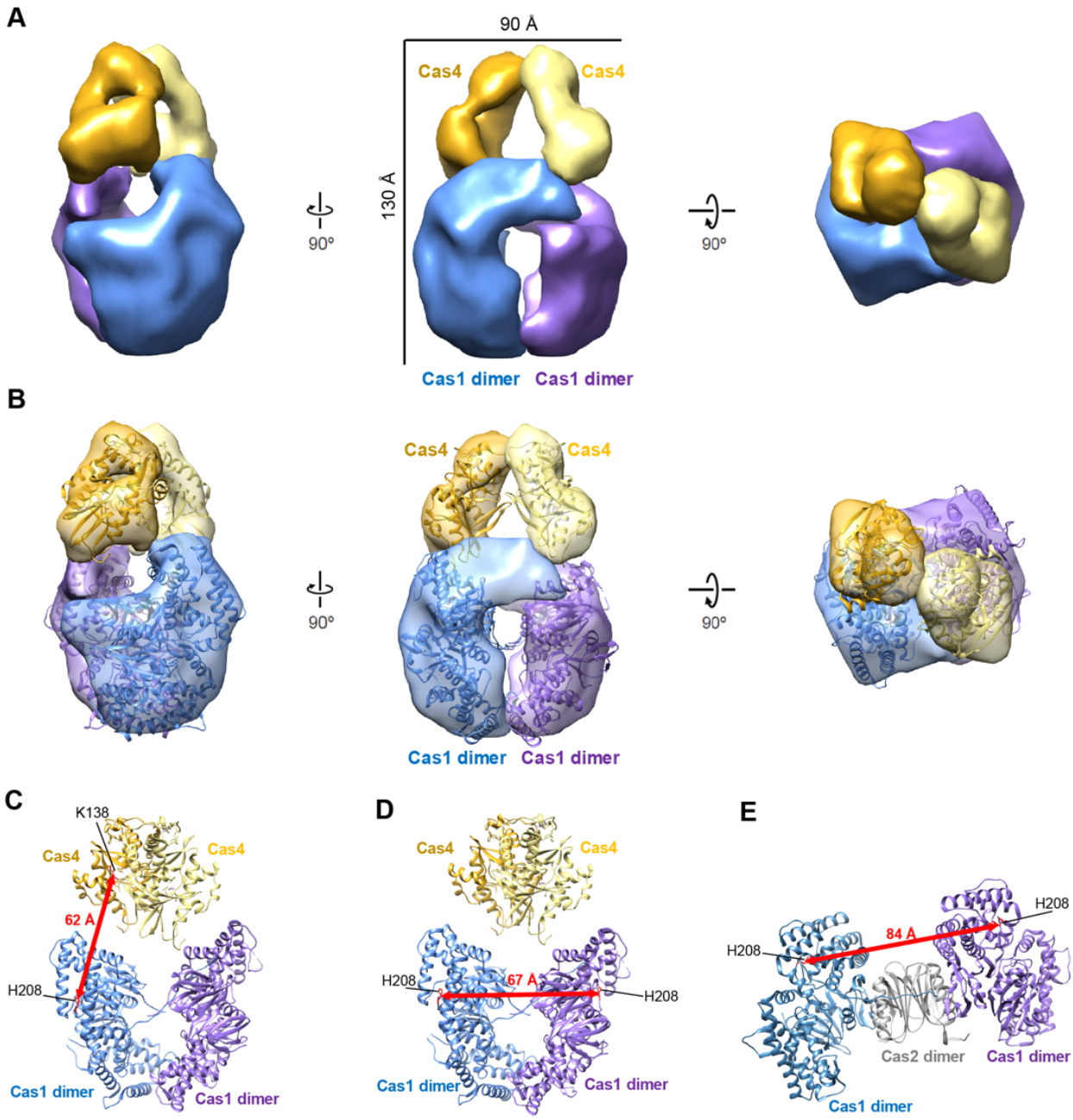

Figure 3. Structure of the *B. halodurans* Cas4-Cas1 complex. (A) Negative-stain reconstruction of the Cas4-Cas1 complex at ~21-Å resolution (using the gold-standard 0.143 FSC criterion) with subunits labeled and colored as follows: gold, Cas4; yellow, second Cas4; blue, Cas1 dimer; purple, second Cas1 dimer. (B) The crystal structures of the Cas4 protein Pcal_0546 from *Pyrobaculum calidifontis* (PDB ID: 4R5Q) and the Cas1 dimers in the Cas1-Cas2 complex from *E.coli* (PDB ID: 4P6I; blue, chain C and D; purple, chain E and F) are docked into the negative-stain reconstruction of the Cas4-Cas1 complex. (C) The distance between the active sites of Cas1 (His234) and Cas4 (Lys110) in the Cas4-Cas1 complex is ~62 Å (red line). (D) The distance between the active sites of two Cas1 (His234) in the Cas4-Cas1 complex is ~67 Å (red line). (E) The distance between the active sites of Cas1 (His234) in *E.coli* Cas1-Cas2 complex is ~84 Å (red line).

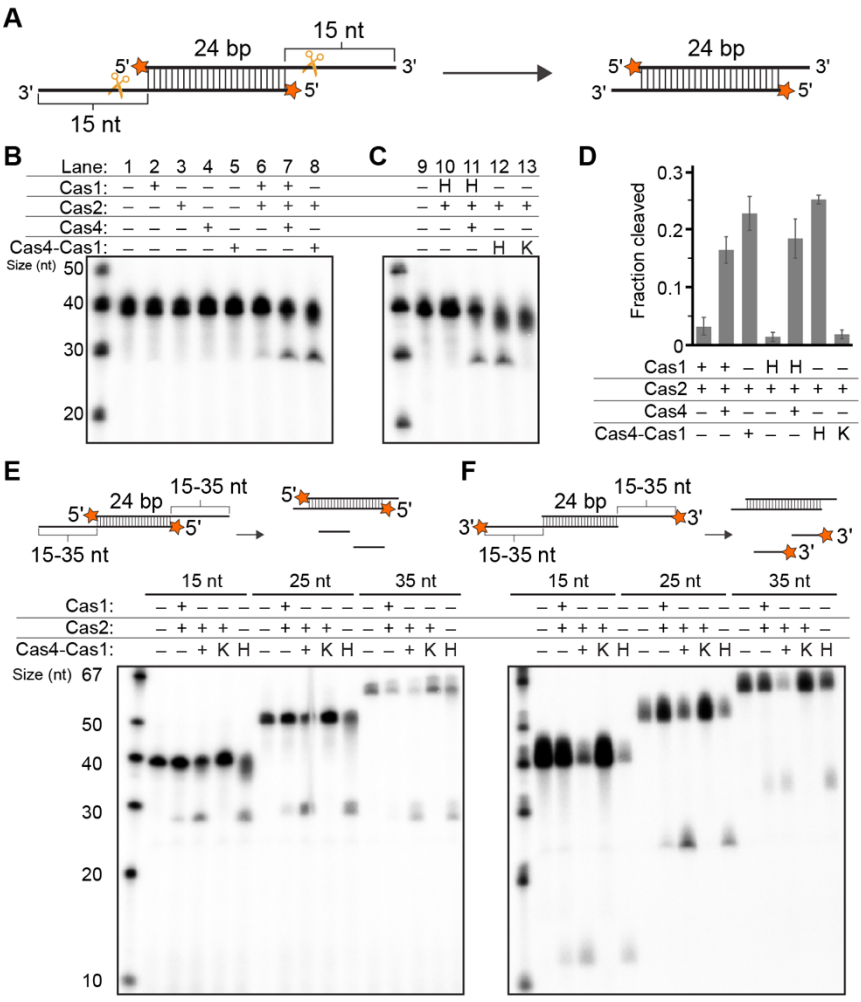

Figure 4. Prespacer processing by the Cas4-Cas1-Cas2 complex. (A) Schematic view of prespacer processing assay. Radiolabel is indicated with a star. (B) Prespacer processing assay using 15 nt 3′-overhang DNA. (C) Prespacer processing assay as in (B) but with catalytic mutants in Cas1 or Cas4 active sites. H indicates H234A Cas1 catalytic mutant; K indicates K110A Cas4 catalytic mutant. (D) Fraction cleaved for prespacer processing. The average of three triplicates is shown, and error bars represent standard deviation. (E-F) Prespacer processing assay with 15 nt, 25 nt, and 35 nt 3′ overhang DNA with radiolabel on (E) 5′ or (F) 3′ end.

Because Cas4 has exonuclease activity, it is possible that the increased processing activity in the presence of Cas4 may be based on exonucleolytic degradation by free Cas4 or Cas4-Cas1. Alternatively, if Cas1-Cas2 engages DNA with overhangs positioned in the Cas4 and/or Cas1 active sites, the cleavage activity may be expected to proceed endonucleolytically. To test these two possibilities, we used substrates with 3′ overhangs of different lengths labeled on either the 5′ or 3′ ends. Interestingly, while all products were the same length for the 5′-end labelled substrates (Fig. 4E), we detected products corresponding

to the length of the overhang for the 3′-end labelled substrates (Fig. 4F), indicating that Cas4 processes prespacers endonucleolytically. Overall, our results suggest that Cas4 is a Cas1-Cas2 dependent endonuclease that processes prespacers at a precise site within the 3′ single-stranded overhang.

PAM-dependent prespacer processing

Given the importance of PAM sequences for targeting by the surveillance complex during CRISPR interference, it is critical that the adaptation complex select and integrate prespacers from sites with correct PAM sequences. In the *B. halodurans* type I-C system, the PAM has been characterized as 5′-GAA-3′ on the target strand (Leenay et al., 2016; Rao et al., 2017) (Fig. 5A). The adaptation complex is expected to recognize the PAM sequence within the 3′-overhang of the prespacer, by analogy to the *E. coli* Cas1-Cas2 complex (Wang et al., 2015). To determine whether Cas4-dependent prespacer processing is sequence specific, we tested the cleavage of two different prespacer substrates containing either 5′-GAA-3′ (perfect) or 5′-TTC-3′ (reverse) PAM on the 3′ overhangs (Fig. 5B). Interestingly, while the presence of Cas4 enhanced the processing of prespacers with a GAA PAM, no detectable cleavage was observed for prespacers with a TTC PAM (Fig. 5C). These data suggest that the adaptation complex can specifically capture prespacers with correct PAM sequences, and that Cas4-dependent cleavage is also PAM dependent.

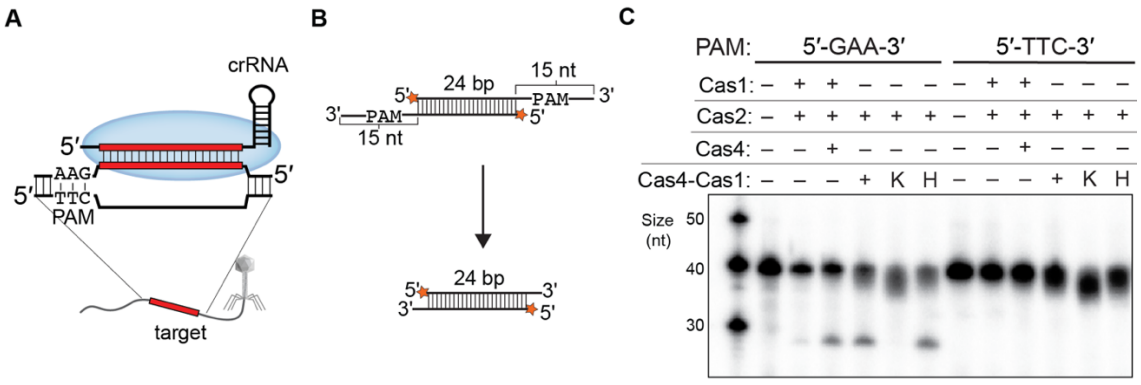

Figure 5. PAM-dependent cleavage by Cas4 in the presence of Cas1-Cas2. (A) Schematic view of PAM sequence (5′-GAA-3′ on the target strand) in the B. halodurans type I-C system. (B) Schematic view of PAM-dependent processing assay. The prespacer contains a PAM site within the 3′ overhang. (C) PAM-dependent processing assay using either 5′-GAA-3′ (perfect) or 5′-TTC-3′ (reverse) PAM on the 3′ overhangs.

Cas4 ensures the integration of processed prespacers

To determine whether Cas4 in the presence of Cas1-Cas2 integrates processed spacers, we designed minimal CRISPRs with varied leader sequence lengths from 10 to 50 bp, the full 32-bp repeat, and a 5-bp spacer. Each minimal CRISPR substrate was labeled at the 5′ end of the plus strand resulting in leader-length products upon successful integration (Fig. 6A). Leader-length products were observed for all minimal CRISPRs (Fig. 6B), indicating that, like the type II-A system, integration by the type I-C adaptation complex relies on intrinsic sequence specificity rather than additional structural motifs or factors, such as IHF (Rollie et al., 2015; Wright and Doudna, 2016; Xiao et al., 2017). Surprisingly, we observe a slight increase in the amount of leader-length product in the presence of Cas4, and especially for the Cas4-Cas1 complex. However, the formation of leader-length products is dependent on Cas1 catalytic activity, as no product is observed with the catalytically dead Cas1 mutant in the absence or presence of Cas4. We also observed the formation of small amounts of leader-length product in the absence of prespacers (Fig. 6C). These data suggest

that Cas1 can perform site-specific cleavage at the leader end, as has been observed previously in the type II-A system (Wright and Doudna, 2016).

Because it is not possible to distinguish between leader-length integration or cleavage products, we also tested integration of preprocessed (5-nt 3′ overhangs) or unprocessed prespacers with varied 3′-overhang lengths in the absence or presence of Cas4. This experimental design ensures the direct detection of integration products, and also enables detection of products containing processed versus unprocessed prespacers (Fig. 7A). Using this experimental design, preprocessed prespacers were integrated with similar efficiency in both the absence and presence of Cas4. These results suggest that Cas4 does not improve the integration efficiency by Cas1 but may contribute to the putative leader cleavage activity observed in the absence of prespacers (Fig. 6C). In contrast, when the unprocessed substrate was used for integration, production of the correct length integration product was substantially enhanced in the presence of Cas4 (Fig. 7B), consistent with the enhanced processing activity by Cas4. Cas1 active site mutants ablated integration activity, whereas Cas4 active site mutants produced integration products only with the preprocessed prespacer, consistent with the lack of processing activity for this mutant (Fig. 7B). Notably, in time-course assays, Cas1-Cas2 integrated unprocessed prespacer but quickly catalyzed the disintegration in favor of integrating processed prespacers (Fig. 7C). However, we were unable to detect integration of unprocessed prespacers in the presence of WT Cas4, although we saw low levels of unprocessed integration for the K110A mutant (Fig. 7C). Together these results indicate that in the presence of Cas4, the Cas1-Cas2 adaptation complex preferentially integrates processed prespacers based on the enhanced processing activity provided by Cas4.

Sequence-specific integration and asymmetric prespacer processing by the adaptation complex

To determine the processing and integration site for HSI products formed following prespacer processing, we developed an integration assay using a plasmid bearing a portion of the native *B. halodurans* CRISPR locus (pCRISPR). Prespacer integration converts a negatively supercoiled plasmid into different plasmid species, such as relaxed or linear forms of the plasmid for successful integration events or plasmid topoisomers for integration followed by disintegration events (Nuñez et al., 2015b; Wright and Doudna, 2016) (Fig. 8A). We detected relaxed or linearized plasmid species when Cas1, Cas2 or Cas4 were incubated alone with pCRISPR, consistent with native integrase activity of Cas1 or nuclease activities of Cas2 and Cas4 (Nuñez et al., 2015b; Nam et al., 2012; Zhang et al., 2012; Lemak et al., 2013). When both Cas1 and Cas2 were added to the integration reaction, we observed robust integration activity where supercoiled plasmids were converted to a linear form at 37°C in both the presence and absence of Cas4 (Fig. 8B). However, at 65°C we observed multiple plasmid topoisomers under both conditions, which were visible as a ladder of slow migrating DNA bands on a post-stained agarose gel, suggesting disintegration is favored at higher temperature.

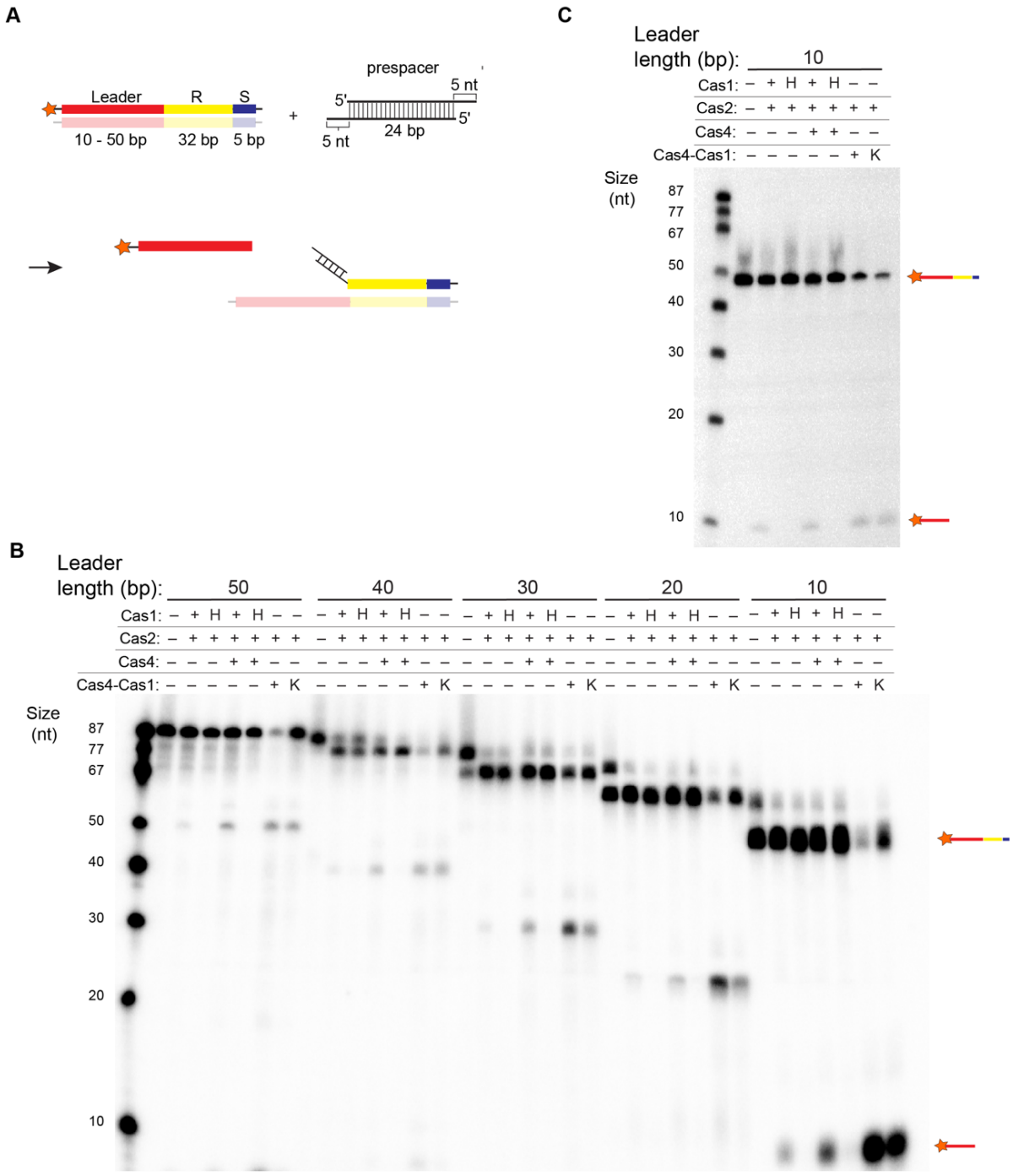

Figure 6. Integration assays using minimal CRISPR. (A) Schematic view of integration assay using a 5′ -radiolabeled minimal CRIAPR with the 5-nt 3′ overhang prespacer. Red, leader; yellow, repeat; blue, spacer; star is radiolabel at indicated position. (B) Integration assay using the short linear CRISPRs with different leader lenghs. (C) Cleavage assay using the minimal CRISPR (10-bp leader) in the absence of prespaer.

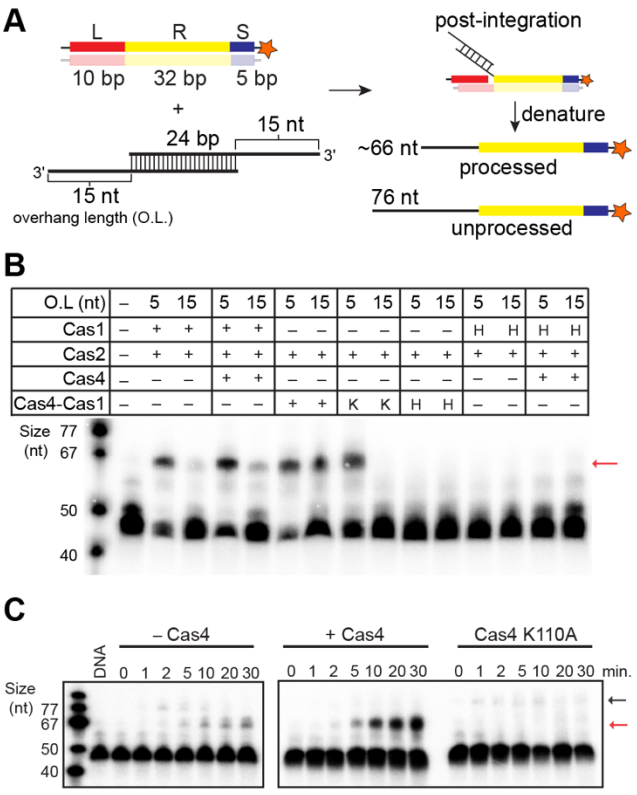

Figure 7. Prespacer processing by Cas4 enhances integration. (A) Schematic view of integration assay. Red, leader; yellow, repeat; blue, spacer; star, radiolabel. The lengths of substrates and expected products are indicated. (B) Integration assay with Cas1 (1 μM), Cas2 (1 μM) and Cas4 (2 μM) individually or Cas4-Cas1 (1 μM) complex containing wild-type (WT) subunits, or Cas1 (H, H234A) or Cas4 active site mutants (K, K110A). Red arrow indicates the integrated products of processed prespacer. O.L is overhang length. (C) Time-course integration assay of WT Cas1 + Cas2, WT Cas4-Cas1 + Cas2, or Cas4-Cas1 + Cas2 with Cas4 active site mutant (K110A) using 15-nt 3′ overhang prespacers. Red arrow indicates the integrated products of processed prespacers while black arrow indicates the integrated products of unprocessed prespacers.

Using this plasmid integration assay, we sought to determine the integration and processing sites for prespacers. Integration events are expected to occur at either the leader-repeat junction of the plus strand or at the first repeat-spacer junction of the minus strand within the CRISPR (Fig. 9A). To determine whether the prespacers were correctly integrated into either the leader-repeat or repeat-spacer junction, we PCR amplified the half-site integrated (HSI) products of Cas1-Cas2 in the presence of Cas4 using a preprocessed prespacer (Fig. 9B). Plus-strand PCR amplicons ran as a single band, while amplification reactions against the minus strand resulted in less specific bands (Fig. 9B), suggesting that plus-strand integration is more specific. We cloned the amplicons into pRSF and sequenced 20 clones for each integration site. All integration events on the plus strand of the CRISPR occurred at the leader-repeat junction, while only 60 % (12 out of 20) of integration events on the minus strand occurred precisely at the repeat-spacer junction (Fig. 9C). These results

indicate that while the majority of integration events occur at the correct site, minus-strand integration is less specific and may be specified following plus-strand integration at the leader-repeat site. Moreover, non-specific minus-strand HSI products may be subject to disintegration, resulting in the formation of topoisomer products at 65°C (Fig. 8B).

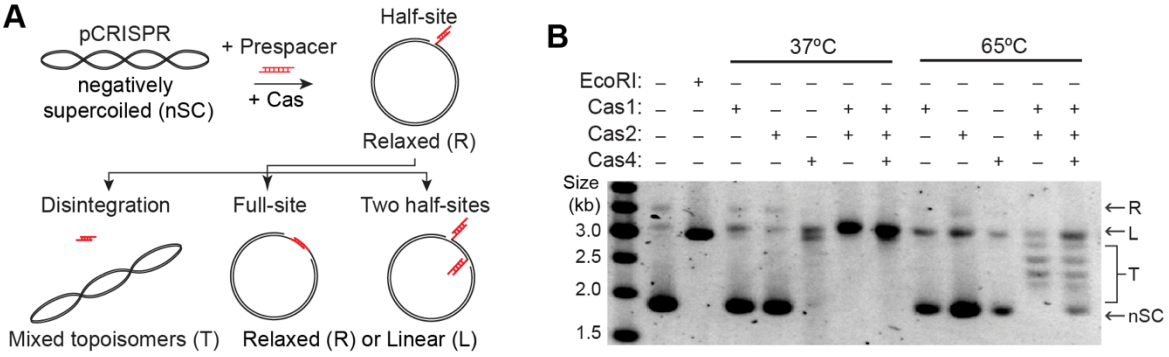

Figure 8. Prespacer integration into pCRISPR. (A) Schematic view of integration assay and possible products using negatively supercoiled pCRISPR plasmid and the pre-processed prespacer. (B) Integration assay using combinations of Cas1, Cas2 and Cas4 with prespacer at 37 °C or 65 °C. The prespacer is a 24-bp duplex flanked by 5-nt 3′ overhangs. EcoRI digested plasmid was used for a linear standard.

We next developed a high throughput sequencing assay with unprocessed prespacers containing degenerate sequences on the 3′ overhangs (Fig. 9A). The degenerate sequences mimic the effects of varied sequences that would be present in prespacers encountered in endogenous situations. These prespacers were used for integration assays into pCRISPR in the presence or absence of Cas4. To limit disintegration of unprocessed HSI products, experiments were performed at 37ºC (Fig. 8B). PCR amplification reactions were performed and the products were sequenced by Illumina MiSeq to determine the integration sites for unprocessed prespacers. Similar to the preprocessed prespacer, the vast majority of HSI products were integrated precisely at the leader-repeat junction in both the presence and absence of Cas4 (Fig. 9D). However, only a small fraction of prespacers were integrated at the expected repeat-spacer junction site on the minus strand (Fig. 9D). The lower specificity of minus strand HSI products for unprocessed prespacers may be due to decreased processing activity observed at 37ºC, resulting in HSI products where the non-integrated overhang

remains unprocessed and unsuitable for full-site integration. Overall, our results suggest that spacer acquisition proceeds through initial integration at the leader-repeat site, and that correct integration at the repeat-spacer site is partially dependent on complete prespacer processing.

Sequencing of HSI products also revealed the extent and site of processing for the prespacer substrates. Consistent with our processing and integration assays (Fig. 3 and 5), the presence of Cas4 greatly enhanced the integration of processed prespacers, although low levels of unprocessed HSI products were detected, potentially because Cas4 was added in trans rather than as part of the Cas4-Cas1 complex (Fig. 9E). Intriguingly, when the degenerate sequence was placed at positions 5-7 of the overhang (Prespacer 1, Fig. 9E), the HSI products displayed a marked asymmetry for the processing sites for each of the prespacer strands. While the “top” strand (green strand, Fig. 9E) was mainly processed following position 4 of the overhang, the “bottom” strand (magenta strand, Fig. 9E) was processed following position 6. For Prespacer 2, in which the degenerate sequence was placed at positions 6-8 of the overhang, the asymmetrical processing sites were also observed, although the processing sites were more variable for this substrate. The variability in processing position for the two different prespacers suggests that sequence specificity plays a role in processing site selection.

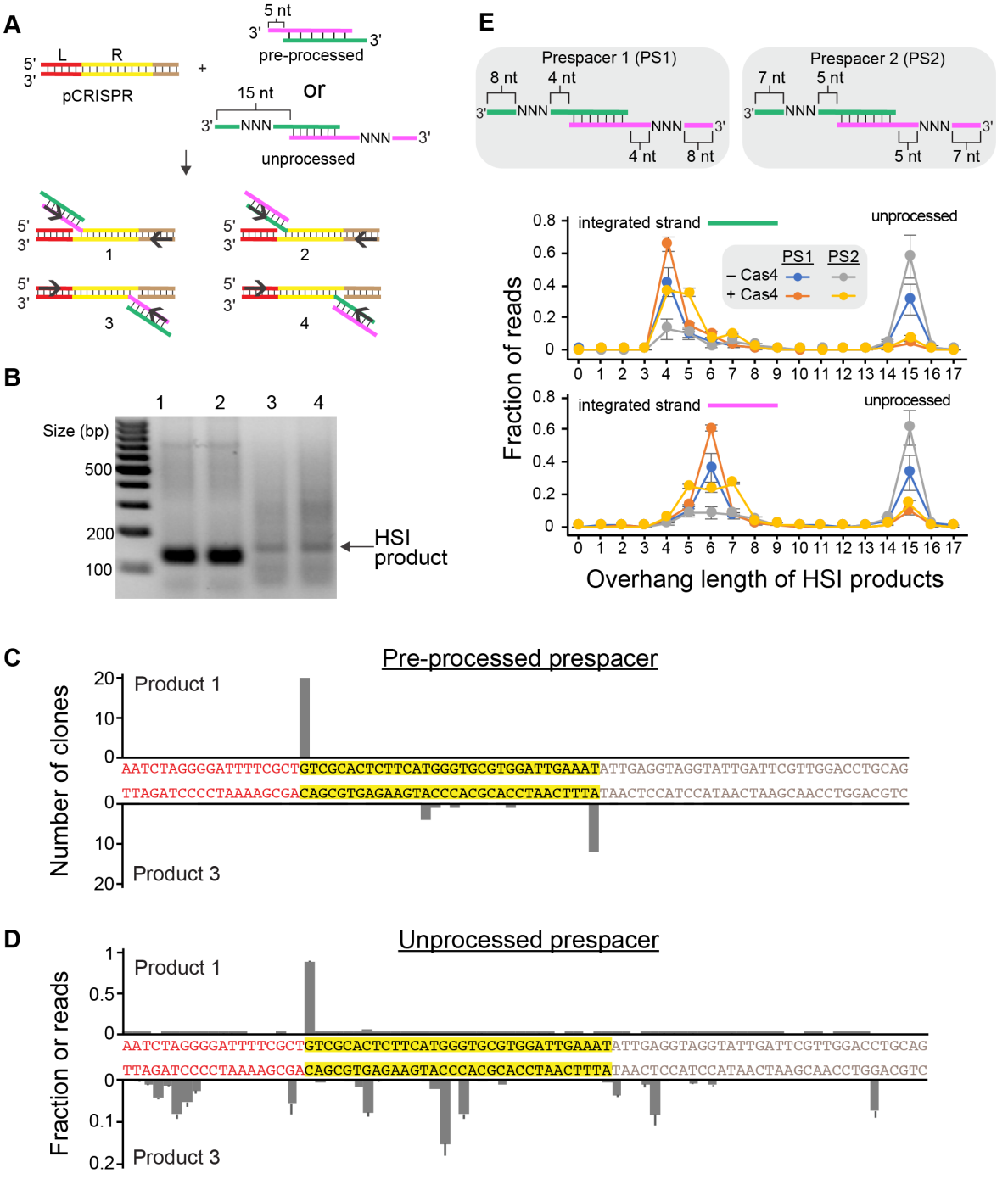

Figure 9. Sequencing half-site products reveal integration and processing sites. (A) Schematic view of half-site integration events. Four different events occurred due to two orientations of prespacers and two different integration sites. Substrates are either preprocessed prespacers or prespacers containing 15 nt 3′ overhangs with degenerate sequences. (B) PCR products for the half-site integrated (HSI) products of Cas1-Cas2 in the presence of Cas4 using a preprocessed prespacer. The numbers indicate the four different events that are depicted in (A). (C) Integration sites for HSI products of Cas1-Cas2 in the presence of Cas4 using the preprocessed spacer. The regions of the CRISPR are colored as in (A). (D) Integration sites for HSI products of Cas1-Cas2 using unprocessed prespacer. The average fraction of read counts at each start site from three separate replicates are plotted, with error bars representing standard deviation. (E) Processing sites for two prespacers for either the top (green) or bottom (magenta) strand in the absence or presence of Cas4. The average fraction of read counts for three separate replicates are plotted, with error bars representing standard deviation.

Discussion

Despite the recognition of Cas4 as a core family of adaptation proteins, it has remained unclear whether it is directly involved in spacer acquisition. Our results reveal that Cas4 is a key factor in ensuring PAM-dependent prespacer processing prior to integration by the Cas1-Cas2 complex. Cas4 is required for efficient processing of precursor prespacers with 3′ overhangs, which may be generated from RecBCD or Cas3 activities (Levy et al., 2015; Kunne et al., 2016; Staals et al., 2016). Previous biochemical studies found that some Cas4 variants exhibit bidirectional exonuclease activity (Lemak et al., 2013), suggesting that Cas4 5′ to 3′ exonucleolytic activity against blunt DNA ends may also generate precursors with 3′ overhangs. A recent study also showed that Cas4 can trim the ends of precursor substrates at multiple sites that are not protected by Cas1-Cas2 to generate the final length of prespacers prior to integration (Rollie et al., 2017). Our results show that Cas4 cuts precursors at specific locations through endonucleolytic cleavage, suggesting that Cas4 associates with Cas1-Cas2 and the complex positions the 3′ overhangs in the Cas4 active site to dictate the exact cut sites.

Our structural studies of the Cas4-Cas1 complex reveal a surprising architecture that may be mutually exclusive with formation of the Cas1-Cas2 complex structure observed in other sub-types. It is possible that type I-C Cas1-Cas2 adopts a different overall conformation. Alternatively, it is possible that the Cas4-Cas1 complex sequesters Cas1, preventing it from forming a productive Cas1-Cas2 complex for integrating dsDNA substrates (Fig. 10A-B). Thus, these competing structures could provide a regulatory mechanism for the adaptation stage of CRISPR immunity. Future structural work will be required to determine how the Cas4-Cas1 structure transitions to the overall adaptation complex.

Our structural studies also indicate that the Cas4 and Cas1 active sites are relatively distal from one another within the Cas4-Cas1 complex. This arrangement of the two active sites disfavors a model in which both proteins simultaneously contribute to cleavage. Consistently, our mutagenesis studies reveal that the Cas4 active site is necessary and sufficient for efficient prespacer processing, while the Cas1 active site can catalyze low levels of cleavage in the absence of Cas4. Together, these results suggest that long single-stranded 3′ overhangs may shuttle between the Cas4 and Cas1 active sites, and may be preferentially bound by Cas4 until cleavage occurs, preventing either cleavage or integration by the Cas1 active site (Fig. 10C-D). Consistently, we mainly observe integration following prespacer processing in the presence of Cas4, while integration of unprocessed prespacers was much more prevalent in the absence of Cas4. In addition, we observe low specificity for minus-strand integration when using unprocessed prespacers, suggesting that specific integration at the repeat-spacer junction only occurs during full-site integration and requires processing of both ends of the substrate (Fig. 10E-F). Integration at the correct repeat-spacer junction is also dictated by specific sequences within the repeat, which provides an additional ruler mechanism to dictate the site of integration (Goren et al., 2016; Wang et al., 2016).

In order to generate functional spacers, prespacers with correct PAMs must be captured by the adaptation complex and processed just upstream of the PAM. Previous studies of the type I-E Cas1-Cas2 complex revealed that Cas1 can recognize and cleave PAM sequences within the 3′-overhangs of prespacers (Wang et al., 2015), while studies of type I-A adaptation indicated that Cas4 trims prespacers in a PAM-dependent manner (Rollie et al., 2017). Similarly, our results show that the type I-C adaptation complex processes prespacers in a PAM-specific manner based on either Cas1-dependent cleavage in the absence of Cas4, or Cas4-dependent cleavage in the presence of all three adaptation proteins. In addition to

enhancing prespacer cleavage, Cas4 also provides an important fidelity check to ensure that prespacers are only integrated following removal of the PAM sequence. Integration prior to processing is likely to result in a spacer targeting a non-PAM region. Thus, our results suggest that Cas4-dependent processing is critical for maintaining a fully functional CRISPR array, without the addition of defective spacers through aberrant integration prior to prespacer processing.

Functional spacers also require that the PAM-proximal end of the prespacer be integrated at the repeat-spacer junction, although it remains unclear how the orientation of spacer integration is achieved. Our results reveal that prespacer processing occurs asymmetrically, with different length overhangs produced following processing. It is possible that asymmetrical overhang length may help to dictate spacer orientation during half-site integration. However, it remains unclear what factors dictate asymmetrical processing. Notably, outside of the degenerate sequence, the overhang sequences were identical for both strands of the prespacers used in our experiments, suggesting that the duplex sequence may affect processing asymmetry rather than the overhang sequence. Further experiments will be needed to determine the exact mechanism for spacer orientation by the adaptation complex.

Our analysis of cleavage site selection also suggests that prespacer length may vary based on “slipping” within the active site during prespacer processing, as processing sites varied for some prespacers tested. These results are consistent with *in vivo* analysis of type I-C spacer acquisition, which has shown that PAM slipping can cause aberrant spacer lengths (Rao et al., 2017). Similarly, the lengths of the existing spacers in the *B. halodurans* CRISPR array varies between 33-36 bp. Together, these data suggest that selection of functional spacers *in vivo* may play a key role in determining the spacer size.

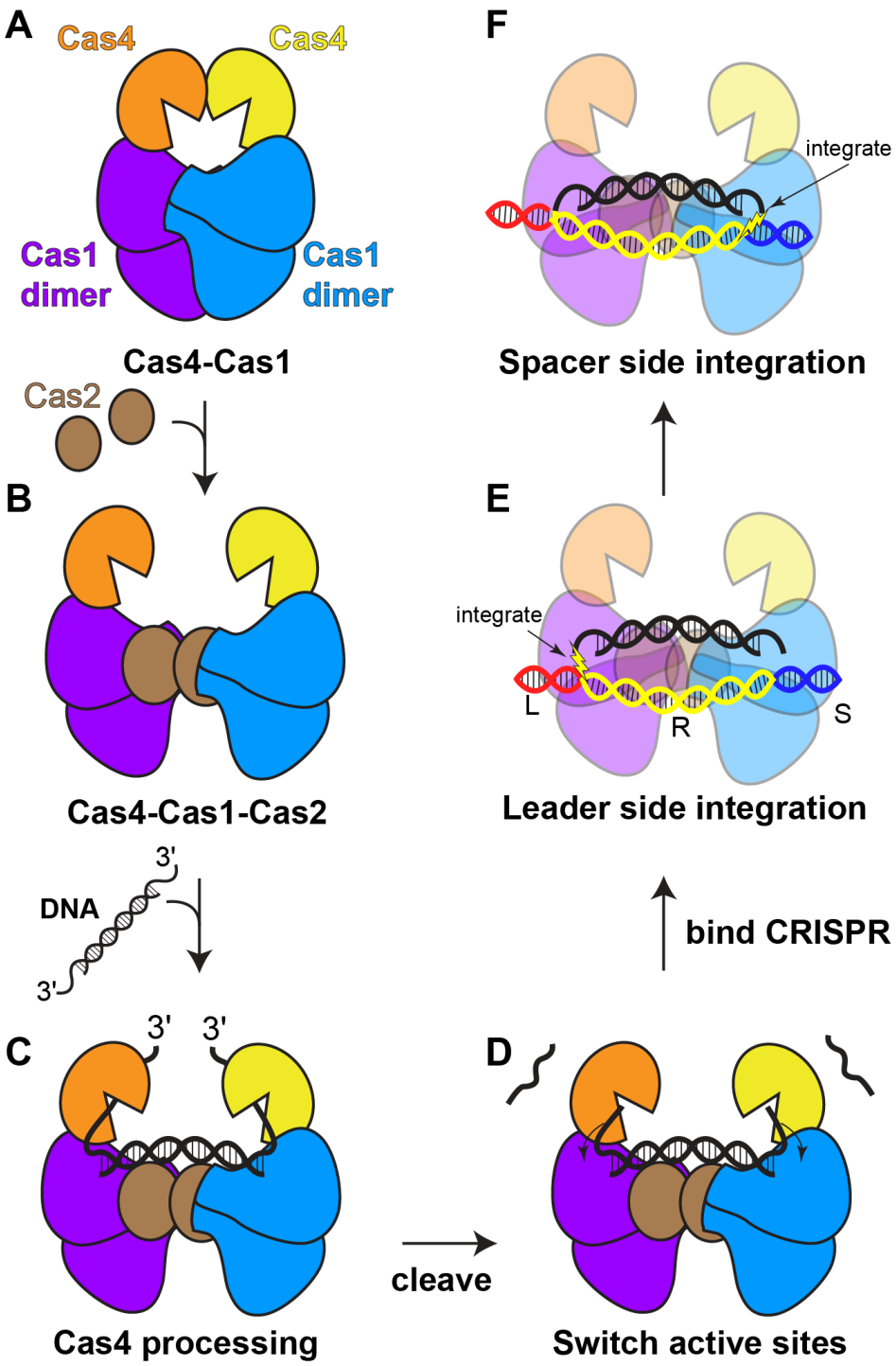

Figure 10. Model of processing and integration by Cas4-Cas1-Cas2. (A-B) Transition of (A) Cas4-Cas1 to (B) a putative Cas4-Cas1-Cas2 complex. The two structures may be mutually exclusive. (C) Upon binding of precursor DNA substrates with long 3′ overhangs, Cas4 subunits within the putative Cas4-Cas1-Cas2 complex binds the overhangs in their active sites. (D) Following cleavage by Cas4, the shortened 3′ overhangs are transferred to the adjacent Cas1 active sites. (E) Following binding of the complex at the CRISPR locus, Cas1 integrates the substrate at the leader-spacer junction. (F) Integration at the repeat-spacer junction is dictated by the length of the substrate and only proceeds following leader-side integration and complete substrate processing.

Overall, our data support a model in which a putative Cas4-Cas1-Cas2 complex controls processing and integration of prespacers (Fig. 10). Cas4 cleavage of precursor prespacers establishes spacer length and PAM site, while active site switching ensures that only cleaved ends enter the Cas1 active site and that integration only occurs following processing at both ends of the precursor (Fig. 10A-B). Precise integration at the leader-repeat junction establishes the location for spacer insertion (Fig. 10E), while integration at the repeat-spacer junction is dictated by the length of the prespacer substrate (Fig. 10F). Our work establishes the core role for Cas4 in type I-C adaptation and suggests a similar integral function in other Cas4-containing systems.

Materials and Methods

Cloning

Genomic DNA from *Bacillus halodurans* was obtained from ATCC. The *cas4*, *cas1*, and *cas2* genes were PCR amplified from the genomic DNA using the indicated primers in Table 1 and cloned into pET52b for *cas4* and pSV272 for *cas2* and *cas1* for individual expression. For co-expression with N-terminal His_6-tagged *cas4*, *cas4* and *cas1* or all three genes were PCR amplified as a single operon and cloned into pET52b. For co-expression with N-terminal His_6-tagged *cas1*, *cas1* was cloned into pET52b and untagged *cas4* was cloned into pRSF. The pET52b expression vector encodes an N-terminal His_6-tag and pSV272 expression vector encodes an N-terminal His_6-MBP (maltose binding protein) tag followed by a tobacco etch virus (TEV) protease recognition site. pCRISPR was generated by PCR amplification of the CRISPR array and leader sequence from the genomic DNA

using the indicated primers in Table 1 and ligation into BamHI- and EcoRI-digested pUC19. All sequences were verified by Sanger sequencing (Eurofins Genomics).

Protein purification

Cas1, Cas2 and Cas4-Cas1 were overexpressed in BL21(DE3) and grown to 0.6 OD_{600} in LB media, followed by overnight induction at 16˚C with 0.5 mM IPTG. The cells were harvested and lysed using a homogenizer (Avestin, Inc.). All proteins were initially purified using HisPur Ni-NTA affinity resin in recommended buffers (Thermo Fisher Scientific). His_6-MBP-Cas1 and His_6-MBP-Cas2 were cleaved using TEV protease overnight at 4˚C to remove His_6-MBP tag. The cleaved Cas1 and Cas2 were flowed through a Ni-NTA column and further purified using a Superdex 200 16/60 or Superdex 75 16/60 GL column (GE Healthcare), respectively, in a size exclusion buffer (20 mM HEPES (pH 7.5), 100mM KCl, 5% glycerol and 2mM DTT).

Cas4 and pRKSUF017 (carrying *sufABCDSE* (Takahashi and Tokumoto, 2002)) were co-expressed in BL21(DE3) star cells and grown to 0.7-0.8 OD_{600} in 2xYT (pH 7.0) media with 100 mg of ferric citrate (Sigma), ferrous sulfate (Fisher), and L-cysteine (MP biomedicals), followed by overnight induction at 18˚C with 1 mM IPTG. Cas4 was purified as described above through a Ni-NTA column and further purified by using Superdex 200 16/600 in size exclusion buffer. All final stocks were concentrated, aliquoted, flash frozen in liquid nitrogen, and stored at -80˚C.

Pull-down assays

32 μM His_6-Cas4-Cas1 or 32 μM His_6-Cas1 was incubated with Ni-NTA SpinTrap columns (Thermo) at 4˚C for 15min in size exclusion buffer. 27 uM or 12.8 μM untagged

Cas2 was added and incubated at 4˚C for 15 min. As a negative control, 27 µM untagged Cas2 was incubated in SpinTrap columns at 4ºC for 15 min in the absence of His_6-Cas4-Cas1. The columns were washed with size exclusion buffer supplemented with 20 mM, 40mM, 80 mM, 150 mM, 200 mM, 250 mM, and 300 mM imidazole, and flow through for each wash was collected for analysis. Samples were run on 4-20% SDS-PAGE (NEB) and the gels were stained by Coomassie blue.

Negative-stain electron microscopy

4 µL Cas4-Cas1 complex (~100 nM) was applied to a glow-discharged copper 400-mesh continuous carbon grid. After one-minute adsorption, the grid was blotted on a filter paper to remove the majority of the protein buffer and immediately stained with 2% (w/v) uranyl acetate solution on the continuous carbon side. The grid was then blotted on a filter paper to remove residual stain and air-dried in a fume hood. The grid was observed with a JEOL 2010F transmission electron microscope operated at 200 keV with a nominal magnification of x60,000 (3.6 Å at the specimen level). Each image was acquired using a 1 s exposure time with a total dose of ~30-35 e-Å-2 and a defocus between –1 and –2 µm. A total of 50 micrographs were manually recorded on a Gatan OneView camera.

Single-particle pre-processing

The image processing and two-dimensional (2D) classification were performed in Appion (Lander et al., 2009). A total of ~21,300 particles were picked from 50 micrographs using a template generated from a 2D class average of another random protein complex with a similar size as the Cas4-Cas1 complex. Particles were extracted using a 64 x 64-pixel box

size. Reference-free 2D class averages were generated using a total of 146 classes. 13,855 particles were left after removal of junk 2D classes in Appion.

Three-dimensional reconstruction and analysis

The 13,855 good particles left after removing those contributing to junk class averages in Appion were first used for ab initio three-dimensional (3D) reconstruction in cryoSPARC (Punjani et al., 2017). These particles were further subjected to reference-free 2D classification with 100 classes in RELION (Scheres, 2012). After further removal of bad 2D classes, 12,967 particles were left for the 3D classification in RELION using the 3D model generated by cryoSPARC as a starting model. The best class (clearest features and with the largest number of particles) was refined using Autorefine within RELION. The reference-free 2D class averages showed excellent agreement with the reprojections of the final 3D model. The Euler angle plot created in RELION showed a good distribution of Euler angles, despite some preferred orientations. The final 3D reconstruction, which showed structural features to ~21 Å resolution based on the 0.143 gold standard FSC criterion, was segmented using Segger (Pintilie et al., 2010) in Chimera (Pettersen et al., 2004). The crystal structures of the Cas4 protein Pcal_0546 from *Pyrobaculum calidifontis* (PDB ID: 4R5Q) (Lemak et al., 2014), Cas1 from *Archaeoglobus fulgidus* (PDB ID: 4N06) (Kim et al., 2013), and the Cas1 dimers in the Cas1-Cas2 complex from *E. coli* (PDB ID: 4P6I; blue, chain C and D; purple, chain E and F) (Nuñez et al., 2014) were docked into the final reconstruction of the Cas4-Cas1 complex. We chose to use the *E. coli* crystal structure for the model shown in Figure 4 because the Cas1 core fits unambiguously into the map. While the *A. fulgidus* variant is a closer homolog to *B. halodurans* Cas1, a large portion of the crystal structure lies outside of the electron density when fit to the map. It appears that the conformation of the

crystal structure is not the same as in our map and would require us to flexibly fit a large alpha-helical domain into the density which would be over-interpreting our structure at the current resolution.

DNA substrate preparation

All oligonucleotides were synthesized by Integrated DNA Technologies. Sequences of all DNA substrates are shown in Table 2. Prespacers and minimal dsDNA CRISPR arrays were hybridized by heating to 95˚C for 5 minutes and slow cooling to room temperature in oligo annealing buffer (20 mM HEPES (pH 7.5), 25 mM KCl, 10 mM $MgCl_2$) and purified on 8% native PAGE. Prespacers were labeled with [γ-^{32}P]-ATP (PerkinElmer) and T4 polynucleotide kinase (NEB) for 5′-end labelling or with [α-^{32}P]-dATP (PerkinElmer) and Terminal Transferase (NEB) for 3′-end labelling. The double-stranded minimal CRISPRs were labeled with [α-^{32}P]-dATP (PerkinElmer) and Klenow-fragment (NEB) for 3′-end labelling.

Integration assays

For plasmid integration assays, Cas4-Cas1 complex was formed by incubating 200 nM Cas1 and 500 nM Cas4 at 37˚C in integration buffer (20 mM HEPES (pH 7.5), 100 mM KCl, 5% glycerol, 2 mM $MnCl_2$, and 2 mM DTT) for 10 min. 200 nM Cas2 was added and the complex was incubated on ice for 10 min. For Cas1-Cas2, 200 nM Cas1 was incubated with 200 nM Cas2 on ice for 10 min. Both Cas1-Cas2 and Cas4-Cas1-Cas2 were incubated with 500 nM of prespacer for 5 min at room temperature. 7.5 nM pCRISPR were added and incubated for one hour either at 37˚C or 65˚C, as indicated. Reactions were quenched by

addition of phenol-chloroform-isoamyl alcohol. Samples were analyzed on 1 % unstained agarose gel at 18 V overnight and post-stained with SYBR Safe (Invitrogen) for 30 min.

Integration assays with 5′- or 3′-radiolabelled minimal CRISPRs were carried out with 2 μM Cas4, 1 μM Cas1, 1 μM Cas4-Cas1, and 1 μM Cas2 in integration buffer. Cas1-Cas2 or Cas4-Cas1-Cas2 complexes were formed from individual protein components through incubation steps described above. The co-purified Cas4-Cas1 complex was incubated with Cas2 at 4ºC for 10 min to form Cas4-Cas1-Cas2. Complexes were incubated with 1 μM prespacer and minimal CRISPRs at 65˚C for 30 minutes. Reactions were quenched by the addition of phenol-chloroform-isoamyl alcohol. Samples were extracted and run on 8% urea-PAGE. The gels were dried and imaged using phosphor screens on a Typhoon imager (GE Life Sciences).

Prespacer processing assays

Prespacer processing assays were performed using 5′or 3′-radiolabelled prespacer with 500 nM Cas4, 200 nM Cas1, 200 nM Cas4-Cas1 and 50 nM Cas2 in integration buffer. The complexes were incubated as described above and all reactions were performed at 65 ˚C for 20 min. Reactions were quenched with phenol-chloroform-isoamyl alcohol. Samples were extracted and were run on 12% urea-PAGE. The gels were dried and imaged using phosphor screens.

Low- and high-throughput sequencing of HSI products

HSI products were amplified from plasmid integration assay products generated as described above, with 2 μM Cas4, 1 μM Cas1 and Cas2, 1 μM of prespacer 1 or 2, and 7.5 nM pCRISPR at 37ºC. For low-throughput sequencing, samples were purified with Wizard

SV Gel and PCR Clean-Up kit (Promega) and amplified using primers (Table 1) containing BamHI or XhoI sites by PCR using GoTaq polymerase (Promega). The amplicons were digested, ligated into BamHI- and XhoI- digested pRSF, and plasmids extracted from 20 transformants for each sample were analyzed by Sanger sequencing. For high-throughput sequencing, three separate samples were prepared for each condition and treated as separate replicates. Samples were purified with PCR clean-up kit before amplification. The samples were amplified by PCR using GoTaq polymerase (Promega) with barcoded primers. Amplification products were analyzed on 2% SYBR Safe stained agarose gels and quantified using densitometry. Samples were mixed in equal quantities and were run on 2% agarose gel. The band was excised and DNA was purified using a Wizard SV Gel and PCR Clean-Up kit (Promega). The DNA was analyzed on a TapeStation 2200 High Sensitivity D1000 kit (Agilent Technologies), libraries were prepared using a TruSeq DNA Nano Library Preparation (Illumina), and libraries were sequenced (2 x 150 paired-end reads) on an Illumina MiSeq by Admera Health, LLC (New Jersey, USA).

HSI Data processing and analysis

Because the vast majority of products were less than 150 bp in length, only the R1 reads were analyzed from the paired-end read output. Sequences were demultiplexed and sorted into separate files for each sample condition and replicate based on the presence of specific pairs of barcodes at both ends of the read using a bash script.

To determine the site of integration, the reads were matched to the pCRISPR sequence using GMAP (Wu and Watanabe, 2005). The site of integration was considered to be the position at which the match between the read and pCRISPR began, with the exception noted below. An output file was generated for each condition and replicate containing the

number of counts at each start site position. The plots in Figure 9 show the average number reads at each start site for the three replicates, with standard deviation represented as error bars. For plus strand HSI products, many prespacer cleavage products ended in T, and were assumed to be integrated at the end of the leader rather than at the -1 position within the leader, which is also a T. Because minus strand integration was less specific, it was not possible to determine both the precise site of integration and the processing site, therefore this assumption was not applied to these products.

To determine the processing site of the positive strand HSI products, the length of sequence between the end of the duplex sequence within the prespacer and the beginning of the repeat sequence within pCRISPR was determined for all reads for amplification products 1 and 2 for all conditions (Fig. 9A). The average number of counts for products of each length for the three replicates is plotted in Figure 9D, with error bars representing standard deviation.

Quantification and statistical analysis

All in vitro experiments were repeated three times, and representative gel images were shown. Quantification of data shown in Fig. 4D was performed using ImageJ. All plotted data are the average of three replicates with error bars representing standard deviation.

Data and software availability

The accession number for the Cas4-Cas1 EM density reported in this paper is EMDB-7485.

Acknowledgements

We thank Raimund Nagel and Reuben Peters for providing pRKSUF017 plasmid and members of the Taylor and Sashital laboratories for helpful discussion. This work was supported by NIH R01 GM115874 (to D.G.S.) and Welch Foundation Grant F-1938 (to D.W.T.). D.W.T is a CPRIT Scholar supported by the Cancer Prevention and Research Institute of Texas (RR160088).

Author Contributions

H.L. performed all biochemical and sequencing experiments. Y.Z. performed single particle electron microscopy and structure determination. H.L., Y.Z., D.W.T., and D.G.S. analyzed and interpreted the results. H.L. and D.G.S. wrote the manuscript with Y.Z. and D.W.T. contributing. D.W.T and D.G.S. supervised research and secured funding for the project.

Declaration of Interests

The authors declare no competing interests.

Table 1. Primers used in this study.

Name	Sequence (5′ → 3′)	Description
1	GTCG GGATCC ATGGCCAGTAATGAAGAAGACCG	Cas4 BamHI forward primer
2	TGTGT CTCGAG TCATTCGCTCAGTCTCCCCTC	Cas4 XhoI reverse primer
3	GTCG GGATCC ATGAAAAAGCTATTAAACACTCTATATGTGAC	Cas1 BamHI forward primer
4	TGTGT CTCGAG CTACTTCCACAGAAATGGCGG	Cas1 XhoI reverse primer
5	GTCG GGATCC ATGCTTGTTTTAATTACGTATGATGTCC	Cas2 BamHI forward primer
6	TGTGT CTCGAG TTAAAAGATAAGAGGGTCCTCTAAATCG	Cas2 XhoI reverse primer
7	TGTGT GAATTC TGGTGCGAACCTCAAGC	pCRISPR EcoRI forward primer
8	GTCG GGATCC GGGTCGGATGATGTCGC	pCRISPR BamH1 reverse primer
9	CGCCATAAAACCGACATAAGCATCAAG	Cas1/Cas4-Cas1 H234A forward primer
10	CAAGACCGTCCTGGCC	Cas1/Cas4-Cas1 H234A reverse primer
11	CGCGTATTCAACAGGAAATGCC	Cas4-Cas1 K110A forward primer
12	CGAGGGAAGCCAAAG	Cas4-Cas1 K110A reverse primer
13[a]	CGTAGCTGAGGACCAC	Forward primer against top strand of prespacer for detecting HSI products
14[a]	CTGTTCTGGTGGTCCTC	Forward primer against bottom strand of prespacer for detecting HSI products
15[a]	GCCAAGCTTGCATGC	Reverse primer against pCRISPR for detecting HSI products integrated in the plus strand
16[a]	ATTCCCTATTTTATCAAAGTGATTTTC	Reverse primer against pCRISPR for detecting HSI products integrated in the minus strand

[a]For HSI product sequencing experiments, restriction enzyme sites or barcodes were added to the 5′-end of primers.

Table 2. Substrate oligonucleotides used in this study.
Bold indicates repeat sequences, RC indicates the complementary strand of the previous strand.

Sequence (5′ → 3′)	Description
CTGTTCTGGTGGTCCTCAGCTACG TTTTG	5 nt 3′-overhang prespacer
CGTAGCTGAGGACCACCAGAACAG TTTTG	RC
AATTCCCTATTTTATCAAAGTGATTTTCTAGAATCTAGGGGATTTTCGCT**GTCGCACTCTTCATGGGTGCGTGGATTGAAAT**ATTGA	50 bp leader
TCAATATTTCAATCCACGCACCCATGAAGAGTGCGACAGCGAAAATCCCCTAGAT TCTAGAAAATCACTTTGATAAAATAGGGAATT	RC
TTTATCAAAGTGATTTTCTAGAATCTAGGGGATTTTCGCT**GTCGCACTCTTCATGGGTGCGTGGATTGAAAT**ATTGA	40 bp leader
TCAATATTTCAATCCACGCACCCATGAAGAGTGCGACAGCGAAAATCCCCTAGAT TCTAGAAAATCACTTTGATAAA	RC
TGATTTTCTAGAATCTAGGGGATTTTCGCT**GTCGCACTCTTCATGGGTGCGTGGATTGAAAT**ATTGA	30 bp leader
TCAATATTTCAATCCACGCACCCATGAAGAGTGCGACAGCGAAAATCCCCTAGATT CTAGAAAATCA	RC
GAATCTAGGGGATTTTCGCT**GTCGCACTCTTCATGGGTGCGTGGATTGAAAT**A TTGA	20 bp leader
TCAATATTTCAATCCACGCACCCATGAAGAGTGCGACAGCGAAAATCCCCTAGATT C	RC
GATTTTCGCT**GTCGCACTCTTCATGGGTGCGTGGATTGAAAT**ATTGA	10 bp leader
TCAATATTTCAATCCACGCACCCATGAAGAGTGCGACAGCGAAAATC	RC
CGTAGCTGAGGACCACCAGAACAG CTCA G	5 nt 3′-overhang prespacer
CTGTTCTGGTGGTCCTCAGCTACG CTCA G	RC
TTTTTTTTAAGTTTT CTGTTCTGGTGGTCCTCAGCTACG	15 nt 5′-overhang prespacer
TTTTTTTTAAGTTTT CGTAGCTGAGGACCACCAGAACAG	RC
CGTAGCTGAGGACCACCAGAACAG TTTTGAATTTTTTTT	Blunt end DNA
AAAAAAAATTCAAAACTGTTCTGGTGGTCCTCAGCTACG	RC
CGTAGCTGAGGACCACCAGAACAG TTTTGAATTTTTTTT	15 nt 3′ overhang prespacer
CTGTTCTGGTGGTCCTCAGCTACG TTTTGAATTTTTTTT	RC
CGTAGCTGAGGACCACCAGAACAG TTTTGAATTTTTTTTTTTTTTTTT	25 nt overhang prespacer
CTGTTCTGGTGGTCCTCAGCTACG TTTTGAATTTTTTTTTTTTTTTTT	RC
CTGTTCTGGTGGTCCTCAGCTACG TTTTGAATTTTTTTTTTTTTTTTTTTTTT	35 nt overhang prespacer
CGTAGCTGAGGACCACCAGAACAG TTTTGAATTTTTTTTTTTTTTTTTTTTTT	RC
CTGTTCTGGTGGTCCTCAGCTACG TTTTGAATTT	10 nt 3′-overhang prespacer
CGTAGCTGAGGACCACCAGAACAG TTTTGAATTT	RC
CGTAGCTGAGGACCACCAGAACAG TTTTGAATTTTTTTTTTTT	20 nt 3′-overhang prespacer
CTGTTCTGGTGGTCCTCAGCTACG TTTTGAATTTTTTTTTTTT	RC
CGTAGCTGAGGACCACCAGAACAG TTTTGAATTTTTTTTTTTTTTTTTTTTTT	30 nt 3′-overhang prespacer
CTGTTCTGGTGGTCCTCAGCTACG TTTTGAATTTTTTTTTTTTTTTTTTTTTT	RC

Table 2. (continued)

CGTAGCTGAGGACCAC CTCA GAA CTGATCGT	16 bp duplex prespacer
GTGGTCCTCAGCTACG CTCA GAA CTGATCGT	RC
CGTAGCTGAGGACCACCAGA CTCA GAA CTGATCGT	20 bp duplex prespacer
TCTGGTGGTCCTCAGCTACG CTCA GAA CTGATCGT	RC
CGTAGCTGAGGACCACCAGAAC CTCA GAA CTGATCGT	22 bp duplex prespacer
GTTCTGGTGGTCCTCAGCTACG CTCA GAA CTGATCGT	RC
CGTAGCTGAGGACCACCAGAACAG CTCA GAA CTGATCGT	24 bp duplex prespacer
CTGTTCTGGTGGTCCTCAGCTACG CTCA GAA CTGATCGT	RC
CGTAGCTGAGGACCACCAGAACAGTA CTCA GAA CTGATCGT	26 bp duplex prespacer
TACTGTTCTGGTGGTCCTCAGCTACG CTCA GAA CTGATCGT	RC
CGTAGCTGAGGACCACCAGAACAGTAGTCG CTCA GAA CTGATCGT	30 bp duplex prespacer
CGACTACTGTTCTGGTGGTCCTCAGCTACG CTCA GAA CTGATCGT	RC
CGTAGCTGAGGACCACCAGAACAGTAGTCGGCTC CTCA GAA CTGATCGT	34 bp duplex prespacer
GAGCCGACTACTGTTCTGGTGGTCCTCAGCTACG CTCA GAA CTGATCGT	RC
GATTTTCGCTGTCGCACTCTTCATGGGTGCGTGGATTGAAATA	10 bp leader for 3′-end labeling
TCAATATTTCAATCCACGCACCCATGAAGAGTGCGACAGCGAAAATC	RC
CGTAGCTGAGGACCACCAGAACAG TTTTNNNTTTTTTTT	NNN for HSI-PS1
CTGTTCTGGTGGTCCTCAGCTACG TTTTNNNTTTTTTTT	RC
CGTAGCTGAGGACCACCAGAACAG TTTTCNNNTTTTTTT	NNN for HSI-PS2
CTGTTCTGGTGGTCCTCAGCTACG TTTTCNNNTTTTTTT	RC
CGTAGCTGAGGACCACCAGAACAG TTTTTTCTTTTTTTT	15 nt 3′ overhang prespacer with TTC PAM
CTGTTCTGGTGGTCCTCAGCTACG TTTTTTCTTTTTTTT	RC

Table 3. Plasmids used in this study.

Plasmids	Description	Primers
Cas4/pET52b	*B. halodurans cas4* expression vector with N-terminal His_6 site tag in pET52b	1 & 2
Cas4-Cas1/pET52b	*B. halodurans cas4-cas1* expression vector with N-terminal His_6 site tag in pET52b	1 & 4
Cas1/pSV272	*B. halodurans cas1* expression vector with N-terminal His_6–MBP-TEV site tag in pSV272	3 & 4
Cas2/pSV272	*B. halodurans cas2* expression vector with N-terminal His_6–MBP-TEV site tag in pSV272	5 & 6
pCRISPR	*B. halodurans* CRISPR locus 4 with leader and one repeat in pUC19	7 & 8

References

Barrangou, R. et al. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. Science *315*, 1709–1712.

Bolotin, A. et al. (2005). Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. Microbiology *151*, 2551–2561.

Brouns, S.J.J. et al. (2008). Small CRISPR RNAs Guide Antiviral Defense in Prokaryotes Stan. Science *321*, 960–964.

Carte, J. et al. (2008). Cas6 is an endoribonuclease that generates guide RNAs for invader defense in prokaryotes. Genes Dev. *22*, 3489–3496.

Deltcheva, E. et al. (2011). CRISPR RNA maturation by trans-encoded small RNA and host factor RNase III. Nature *471*, 602–607.

Fagerlund, R.D. et al. (2017). Spacer capture and integration by a type I-F Cas1–Cas2-3 CRISPR adaptation complex. Proc. Natl. Acad. Sci. 201618421.

Garneau, J.E. et al. (2010). The CRISPR/Cas bacterial immune system cleaves bacteriophage and plasmid DNA. Nature *468*, 67–71.

Gesner, E.M. et al. (2011). Recognition and maturation of effector RNAs in a CRISPR interference pathway. Nat. Struct. Mol. Biol. *18*, 688–692.

Goren, M.G. et al. (2016). Repeat Size Determination by Two Molecular Rulers in the Type I-E CRISPR Array. Cell Rep. *16*, 2811–2818.

Hatoum-Aslan, A. et al. (2013). A ruler protein in a complex for antiviral defense determines the length of small interfering CRISPR RNAs. J. Biol. Chem. *288*, 27888–27897.

Haurwitz, R.E. et al. (2010). Sequence- and structure-specific RNA processing by a CRISPR endonuclease. Science *329*, 1355–1358.

Hochstrasser, M.L., and Doudna, J.A. (2015). Cutting it close: CRISPR-associated endoribonuclease structure and function. Trends Biochem. Sci. *40*, 58–66.

Hudaiberdiev, S. et al. (2017). Phylogenomics of Cas4 family nucleases. BMC Evol. Biol. *17*, 232.

Jackson, R.N., and Wiedenheft, B. (2015). A Conserved Structural Chassis for Mounting Versatile CRISPR RNA-Guided Immune Responses. Mol. Cell *58*, 722–728.

Jackson, S.A. et al. (2017). CRISPR-Cas: Adapting to change. Science *356*.

Kim, T.Y. et al. (2013). Crystal structure of Cas1 from Archaeoglobus fulgidus and characterization of its nucleolytic activity. Biochem. Biophys. Res. Commun. *441*, 720–725.

Koonin, E. V. et al. (2017). Diversity, classification and evolution of CRISPR-Cas systems. Curr. Opin. Microbiol. *37*, 67–78.

Kunne, T. et al. (2016). Cas3-Derived Target DNA Degradation Fragments Fuel Primed CRISPR Adaptation. Mol. Cell 1–13.

Lander, G.C. et al. (2009). Image Processing. Access *166*, 95–102.

Leenay, R.T. et al. (2016). Identifying and Visualizing Functional PAM Diversity across CRISPR-Cas Systems. Mol. Cell *62*, 137–147.

Lemak, S. et al. (2013). Toroidal structure and DNA cleavage by the CRISPR-associated [4Fe-4S] cluster containing Cas4 nuclease SSO0001 from sulfolobus solfataricus. J. Am. Chem. Soc. *135*, 17476–17487.

Lemak, S. et al. (2014). The CRISPR-associated Cas4 protein Pcal 0546 from Pyrobaculum calidifontis contains a [2Fe-2S] cluster : crystal structure and nuclease activity. Nucleic Acids Res. *42*, 11144–11155.

Levy, A. et al. (2015). CRISPR adaptation biases explain preference for acquisition of foreign DNA. Nature *520*, 505–510.

Li, M. et al. (2014). Adaptation of the Haloarcula hispanica CRISPR-Cas system to a purified virus strictly requires a priming process. Nucleic Acids Res. *42*, 2483–2492.

Marraffini, L.A. (2015). CRISPR-Cas immunity in prokaryotes. Nature *526*, 55–61.

Marraffini, L.A., and Sontheimer, E.J. (2008). CRISPR Interference Limits Horizontal Gene Transfer in Staphylococci by Targeting DNA. Science *322*, 1843–1845.

McGinn, J., and Marraffini, L.A. (2016). CRISPR-Cas Systems Optimize Their Immune Response by Specifying the Site of Spacer Integration. Mol. Cell *64*, 616–623.

Mohanraju, P. et al. (2016). Diverse evolutionary roots and mechanistic variations of the CRISPR-Cas systems. Science *353*, aad5147.

Mojica, F.J.M. et al. (2005). Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. J. Mol. Evol. *60*, 174–182.

Mojica, F.J.M. et al. (2009). Short motif sequences determine the targets of the prokaryotic CRISPR defence system. Microbiology *155*, 733–740.

Nuñez, J.K. et al. (2014). Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. Nat. Struct. Mol. Biol. *21*, 528–534.

Nuñez, J.K. et al. (2015a). Foreign DNA capture during CRISPR–Cas adaptive immunity. Nature *527*, 535–538.

Nuñez, J.K. et al. (2015b). Integrase-mediated spacer acquisition during CRISPR–Cas adaptive immunity. Nature *519*, 193–198.

Nuñez, J.K. et al. (2016). CRISPR Immunological Memory Requires a Host Factor for Specificity. Mol. Cell *62*, 824–833.

Pettersen, E.F. et al. (2004). UCSF Chimera - A visualization system for exploratory research and analysis. J. Comput. Chem. *25*, 1605–1612.

Pintilie, G.D. et al. (2010). Quantitative analysis of cryo-EM density map segmentation by watershed and scale-space filtering, and fitting of structures by alignment to regions. J. Struct. Biol. *170*, 427–438.

Plagens, A. et al. (2012). Characterization of the CRISPR/Cas subtype I-A system of the hyperthermophilic crenarchaeon Thermoproteus tenax. J. Bacteriol. *194*, 2491–2500.

Pourcel, C. et al. (2005). CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. Microbiology *151*, 653–663.

Punjani, A. et al. (2017). CryoSPARC: Algorithms for rapid unsupervised cryo-EM structure determination. Nat. Methods *14*, 290–296.

Rao, C. et al. (2017). Priming in a permissive type I-C CRISPR-Cas system reveals distinct dynamics of spacer acquisition and loss. RNA *23*, 1525–1538.

Redding, S. et al. (2015). Surveillance and Processing of Foreign DNA by the Escherichia coli CRISPR-Cas System. Cell.

Rollie, C. et al. (2015). Intrinsic sequence specificity of the Cas1 integrase directs new spacer acquisition. Elife *4*.

Rollie, C. et al. (2017). Prespacer processing and specific integration in a Type I-A CRISPR system. Nucleic Acids Res. 1–14.

Rollins, M.F. et al. (2017). Cas1 and the Csy complex are opposing regulators of Cas2/3 nuclease activity. Proc. Natl. Acad. Sci. *1*, 201616395.

Sashital, D.G. et al. (2011). An RNA-induced conformational change required for CRISPR RNA cleavage by the endoribonuclease Cse3. Nat. Struct. Mol. Biol. *18*, 680–687.

Scheres, S.H.W. (2012). RELION: Implementation of a Bayesian approach to cryo-EM structure determination. J. Struct. Biol. *180*, 519–530.

Semenova, E. et al. (2011). Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. Proc. Natl. Acad. Sci. *108*, 10098–10103.

Staals, R.H.J. et al. (2016). Interference-driven spacer acquisition is dominant over naive and primed adaptation in a native CRISPR–Cas system. Nat. Commun. *7*, 1–13.

Sternberg, S.H. et al. (2014). DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. Nature *507*, 62–67.

Takahashi, Y., and Tokumoto, U. (2002). A third bacterial system for the assembly of iron-sulfur clusters with homologs in archaea and plastids. J. Biol. Chem. *277*, 28380–28383.

Wang, J. et al. (2015). Structural and Mechanistic Basis of PAM-Dependent Spacer Acquisition in CRISPR-Cas Systems. Cell *163*, 840–853.

Wang, R. et al. (2016). DNA motifs determining the accuracy of repeat duplication during CRISPR adaptation in Haloarcula hispanica. Nucleic Acids Res. *44*, 4266–4277.

Westra, E.R. et al. (2012). CRISPR Immunity Relies on the Consecutive Binding and Degradation of Negatively Supercoiled Invader DNA by Cascade and Cas3. Mol. Cell *46*, 595–605.

Wright, A. V. et al. (2017). Structures of the CRISPR genome integration complex. Science *357*, 1113–1118.

Wright, A. V., and Doudna, J.A. (2016). Protecting genome integrity during CRISPR immune adaptation. Nat. Struct. Mol. Biol. *23*, 876–883.

Wu, T.D., and Watanabe, C.K. (2005). GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics *21*, 1859–1875.

Xiao, Y. et al. (2017). How type II CRISPR–Cas establish immunity through Cas1–Cas2-mediated spacer integration. Nature *550*, 137–141.

Xue, C. et al. (2017). Real-Time Observation of Target Search by the CRISPR Surveillance Complex Cascade. Cell Rep. *21*.

Yosef, I. et al. (2012). Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. Nucleic Acids Res. *40*, 5569–5576.

Zhang, J. et al. (2012). The CRISPR Associated Protein Cas4 Is a 5′ to 3′ DNA Exonuclease with an Iron-Sulfur Cluster. PLoS One *7*.

CHAPTER 3. CAS4-CAS1-CAS2 CRISPR ADAPTATION COMPLEX PROCESSES SINGLE STRAND DNA SEQUENCE SPECIFICALLY

Introduction

Bacteria and archaea use an adaptive immune system composed of clustered regularly interspaced short palindromic repeats (CRISPR) arrays and CRISPR-associated (Cas) proteins to defend against infection (Barrangou et al., 2007; Brouns et al., 2008; Marraffini and Sontheimer, 2008). Within this system, the CRISPR locus is programmed with "spacer" sequences that are derived from foreign DNA and serve as a record of prior infection events (Bolotin et al., 2005; Mojica et al., 2005; Pourcel et al., 2005). The host adapts to an infection event when Cas proteins insert short fragments from the invader DNA as new spacers between repeating sequence elements within the CRISPR locus (reviewed in Jackson et al., 2017). The locus is transcribed and processed into short CRISPR RNAs (crRNAs), which assemble with Cas proteins to form a RNA-guided surveillance complex (reviewed in Hochstrasser and Doudna, 2015; Charpentier et al., 2015). Finally, the surveillance complex recognizes the target bearing complementarity to the crRNA sequence and a Cas nuclease cleaves or degrades the target during the interference stage (reviewed in Marraffini, 2015)

Although the machinery and mechanisms involved in CRISPR interference are extremely diverse (Koonin et al., 2017), the adaptation proteins Cas1 and Cas2 are conserved among most CRISPR systems, suggesting a common molecular mechanism for acquiring spacers. Cas1 and Cas2 form a heterohexameric complex that catalyzes spacer integration via two transesterification reactions mediated by nucleophilic attack of the 3′-hydroxyl on each strand of a double-stranded prespacer substrate at the phosphodiester backbone within the CRISPR array. Integration occurs at the first repeat in the CRISPR array, with one attack occurring between the upstream leader sequence and the repeat and the other occurring on the opposite strand between the repeat and first spacer within the array (Nuñez et al., 2015b;

Rollie et al., 2015). These reactions result in the insertion of the prespacer between two single-strand repeats, and this gapped intermediate is repaired by host factors (Ivančić-Baće et al., 2015).

In order to form a functional spacer, the adaptation complex must capture and process longer fragments of DNA from the invader containing a flanking sequence called a protospacer adjacent motif, PAM (Nuñez et al., 2015a; Wang et al., 2015; Xiao et al., 2017). The PAM is an essential motif during target recognition by the surveillance complex and must be present next to the target in order for interference to occur (Deveau et al., 2008; Semenova et al., 2011; Sashital et al., 2012; Sternberg et al., 2014; Redding et al., 2015). However, the PAM is not part of the spacer and must be removed from the prespacer prior to integration through a processing step. In addition, integration must occur in the correct orientation to produce a crRNA that is complementary the PAM-containing strand of the invader.

In some systems, additional *cas* genes, such as Cas4, are also required during adaptation. Cas4 is widespread in type I, II, and V systems (Hudaiberdiev et al., 2017). In *in vivo* studies, deletion of *cas4* reduced the adaptation efficiency (Li et al., 2014) and resulted in the acquisition of non-functional spacers from regions that lacked a correct PAM (Kieper et al., 2018; Shiimori et al., 2018). Some systems have two *cas4* genes that work together to define the PAM, length and orientation of spacers, suggesting that the two Cas4 proteins are involved in processing each end of the prespacer and that they may be present during integration (Shiimori et al., 2018). Similarly, *in vitro* studies have suggested that Cas4 is involved in PAM-dependent prespacer processing (Lee et al., 2018; Rollie et al., 2017). Cas4 endonucleolytically cleaves PAM-containing 3′-single-stranded overhangs that flank double-stranded prespacers (Lee et al., 2018). Importantly, Cas4 cleavage activity is dependent on

the presence of Cas1 and Cas2, and Cas4 inhibits premature integration of unprocessed prespacers. These observations suggest that Cas4 associates with Cas1-Cas2 complex, although direct biochemical and structural evidence for this Cas4-Cas1-Cas2 complex remains elusive (Plagens et al., 2012; Lee et al., 2018).

Here we show that Cas4 forms a complex with Cas1-Cas2 in the presence of CRISPR DNA. Using single-particle negative-stain electron microscopy (EM), we determined the architecture of *Bacillus halodurans* type I-C Cas1-Cas2 and Cas4-Cas1-Cas2 complexes. Unlike the symmetrical $Cas1_4$-$Cas2_2$ structure, we observed a mixture of symmetrical ($Cas4_2$-$Cas1_4$-$Cas2_2$) and asymmetrical ($Cas4_1$-$Cas1_4$-$Cas2_2$) complexes, suggesting a structural mechanism for distinguishing between the PAM and non-PAM end of the prespacer following processing. Surprisingly, the Cas4-Cas1-Cas2 complex processes single-strand DNA when an activator duplex DNA is provided *in trans*. Using this ssDNA cleavage assay, we show that the Cas4-Cas1-Cas2 complex is highly specific for PAM sequences and cleaves precisely upstream of the PAM. Collectively, these findings provide the first structural information of the Cas4-Cas1-Cas2 adaptation complex and reveal the precision and specificity of prespacer processing prior to integration.

Results

Formation of the Cas4-Cas1-Cas2 complex

We previously showed that *B. halodurans* type I-C Cas4 associates tightly with Cas1 but were unable to obtain the Cas4-Cas1-Cas2 complex due to instability of the Cas1-Cas2 complex in this system (Lee et al., 2018). We hypothesized that CRISPR DNA substrates may help stabilize the complex. To test this possibility, we designed a hairpin target containing a 10-bp leader, the full 32-bp repeat, and a 5-bp spacer, mimicking the region of

the CRISPR at which integration occurs (Fig. 1A). We incubated the individually purified Cas1 and Cas2 proteins with hairpin target DNA in equimolar amounts and removed unassociated DNA via ion-exchange chromatography followed by size-exclusion chromatography to remove free Cas proteins. Incubation of individual components with or without the hairpin target led to different elution volumes from a size-exclusion column. In the absence of the target, Cas1 and Cas2 proteins generated two separate peaks, which correspond to the peaks of each individual component (Fig. 1C). When Cas1 and Cas2 were incubated with the target DNA, the proteins eluted earlier as a single peak, while unassociated Cas2 eluted at the original elution volume (Fig. 1B-C). These data indicate that the Cas1-Cas2 complex from type I-C is stabilized in the presence of dsDNA.

Next, we attempted to reconstitute the putative Cas4-Cas1-Cas2 complex in the presence of the CRISPR hairpin target (Fig. 1A). Following incubation of equimolar amounts of each component, Cas4, Cas1 and Cas2 eluted in a single peak from the size-exclusion column, with an earlier elution volume than the Cas1-Cas2-target sample (Fig. 1B-C). In addition, we observed two peaks with the approximate elution volumes of free Cas1 and Cas2 (Fig. 2). Both complexes contained the hairpin target DNA (Fig. 1D). Together, these data suggest that the proteins directly interact with the hairpin target, and the formation of the higher-order complex is also stabilized by the presence of dsDNA substrates.

Architecture of the Cas4-Cas1-Cas2 complex

To characterize the molecular architecture of the complexes, we next performed single-particle electron microscopy (EM) of negatively stained Cas1-Cas2 or Cas4-Cas1-Cas2 complexes bound to the target (Fig. 1E-F, 3-4). Raw micrographs and two-

dimensional (2D) class averages revealed particles with fairly homogenous size and symmetrical architecture consistent with the known structure of the Cas1-Cas2 complex (Nuñez et al., 2014, 2015b; Wang et al., 2015; Xiao et al., 2017a) (Fig. 4A-B). Some 2D class averages for the Cas4-Cas1-Cas2 complex contained clear additional density, suggesting the presence of ordered Cas4 within the complex (Fig. 1E-F).

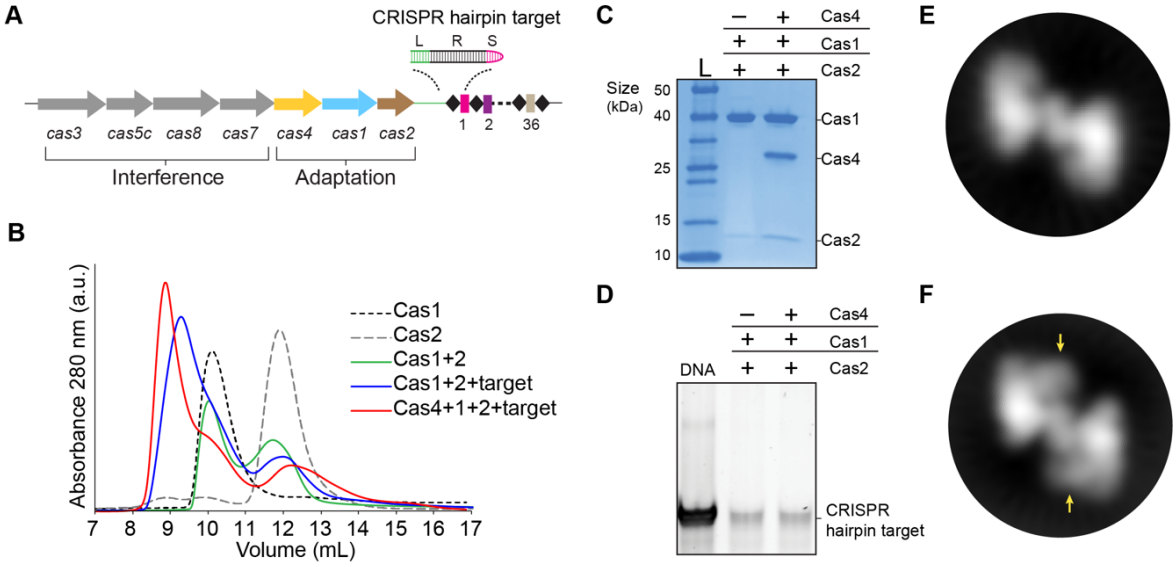

Figure 1. Complex formation of *B. halodurans* Cas1-Cas2 or Cas4-Cas1-Cas2 in the presence of CRISPR hairpin target. (A) Overview of the type I-C *cas* genes and CRISPR locus found in the *Bacillus halodurans* type I-C system. Spacers are shown as rectangles, repeats are shown as diamonds, each *cas* gene is shown as an arrow and gene products involved in adaptation or interference are indicated. The hairpin targets used for this study contains a 10-bp leader, the full 32-bp repeat, and a 5-bp spacer. (B) Size- exclusion chromatography (SEC) of Cas1-Cas2 bound to the targets Cas4-Cas1-Cas2 bound to the targets with individually purified Cas1, Cas2, and Cas4. Cas1-Cas2 without the targets eluted separately on the column. (C) Coomassie-blue stained SDS/PAGE gel of purified complexes. (D) SYBR Gold stained 10% PAGE gel (E) Representative 2D class average of the Cas1-Cas2 complex. (F) Representative 2D class average of the Cas4-Cas1-Cas2 complex. Extra density corresponding to Cas4 is indicated by arrows.

For the Cas1-Cas2 complex, we determined a 19 Å three-dimensional (3D) reconstruction (Fig. 3A). The EM density revealed clear C2 symmetry, which was enforced during the final 3D refinement (Fig. 5A). Segmentation of the density revealed three clear domains, corresponding to two Cas1 dimers sandwiching a Cas2 dimer. The crystal structure of the type II-A Cas1-Cas2-prespacer complex from *Enterococcus faecalis* (Fig. 3A) fit in the density better than a similar structure from the *E. coli* type I-E Cas1-Cas2 complex (Fig.

3B), revealing that the architecture of type I-C Cas1-Cas2 may be more similar to type II-A than to another type I system. Due to the fact that uranyl formate does not stain nucleic acid well (Nogales and Scheres, 2015) we were unable to assign density for the DNA component of the Cas1-Cas2-target complex.

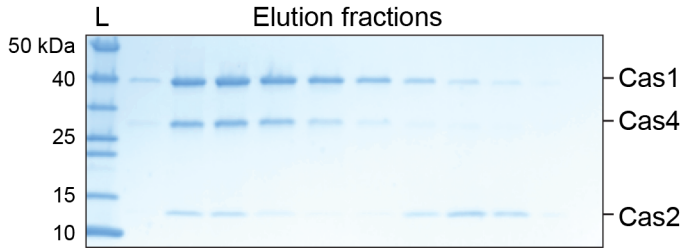

Figure 2. Coomassie-blue stained SDS/PAGE gel of the fractions eluted from the size exclusion column when purifying the complex of Cas4-Cas1-Cas2 with the targets.

For the Cas4-Cas1-Cas2 complex, we determined a 16 Å 3D reconstruction of symmetrical particles enforcing C2 symmetry (Fig. 5B). The segmented density clearly reveals the base Cas1-Cas2 architecture, along with additional density corresponding to two molecules of Cas4 (Fig. 3C). During 3D classification of particles, we observed two classes containing approximately 50% of Cas4-Cas1-Cas2 particles that appeared to contain density for only a single Cas4 molecule (Fig. 5B). A subset of these particles was refined as a separate 3D reconstruction without symmetry enforced, revealing an asymmetrical Cas4-Cas1-Cas2 complex with 1:4:2 stoichiometry (Fig. 3D). These particles may represent ternary complexes in a partially dissociated state, or incomplete formation of the 2:4:2 stoichiometry complex due to reduced affinity for Cas4 within the ternary complex.

In both the symmetrical and asymmetrical Cas4-Cas1-Cas2 reconstructions, the Cas4 Cas1 density is contiguous but Cas4 appears to be distinct from the Cas2 density (Fig. 3C-D). This suggests that Cas4 interacts with Cas1 but not Cas2 within the ternary complex, consistent with the tight interaction we have previously observed between Cas4 and Cas1 in the absence of Cas2 (Lee et al., 2018). However, the interaction surface between Cas4-Cas1

appears to be different in the context of the binary and ternary complexes, suggesting that Cas4 and Cas1 may have two alternative modes of interaction (Fig. 1E-F). We modeled the crystal structure of Cas4 from *Pyrobaculum calidifontis* (PDB: 4R5Q) into the Cas4 density for each Cas4-Cas1-Cas2 reconstruction (Fig. 1C-D). The segmented density for Cas4 is smaller than the structure, suggesting that the segmentation between Cas1 and Cas4 densities was incomplete due to low resolution. Although we cannot confidently assign the orientation of Cas4 within this density, we note that the strongest electron density within the assigned Cas4 volume may correspond to an electron-dense Fe_2-S_2 cluster bound by Cas4 (Fig. 1E). Modelling Cas4 with the Fe_2-S_2 cluster contained within this strong density positions the Cas4 active site in closer proximity to the Cas1 active site.

Cas4 is activated for ssDNA processing in the presence of dsDNA

The observation that the presence of CRISPR DNA substrates stabilized the Cas4-Cas1-Cas2 complex prompted us to hypothesize that binding to the CRISPR DNA may enhance processing activity by stimulating complex formation. To test this hypothesis, we tested cleavage of a prespacer substrate used in our previous study (Lee et al., 2018), containing a 24-bp duplex with 15-nt 3overhangs in the absence of CRISPR DNA (Fig. 6A).

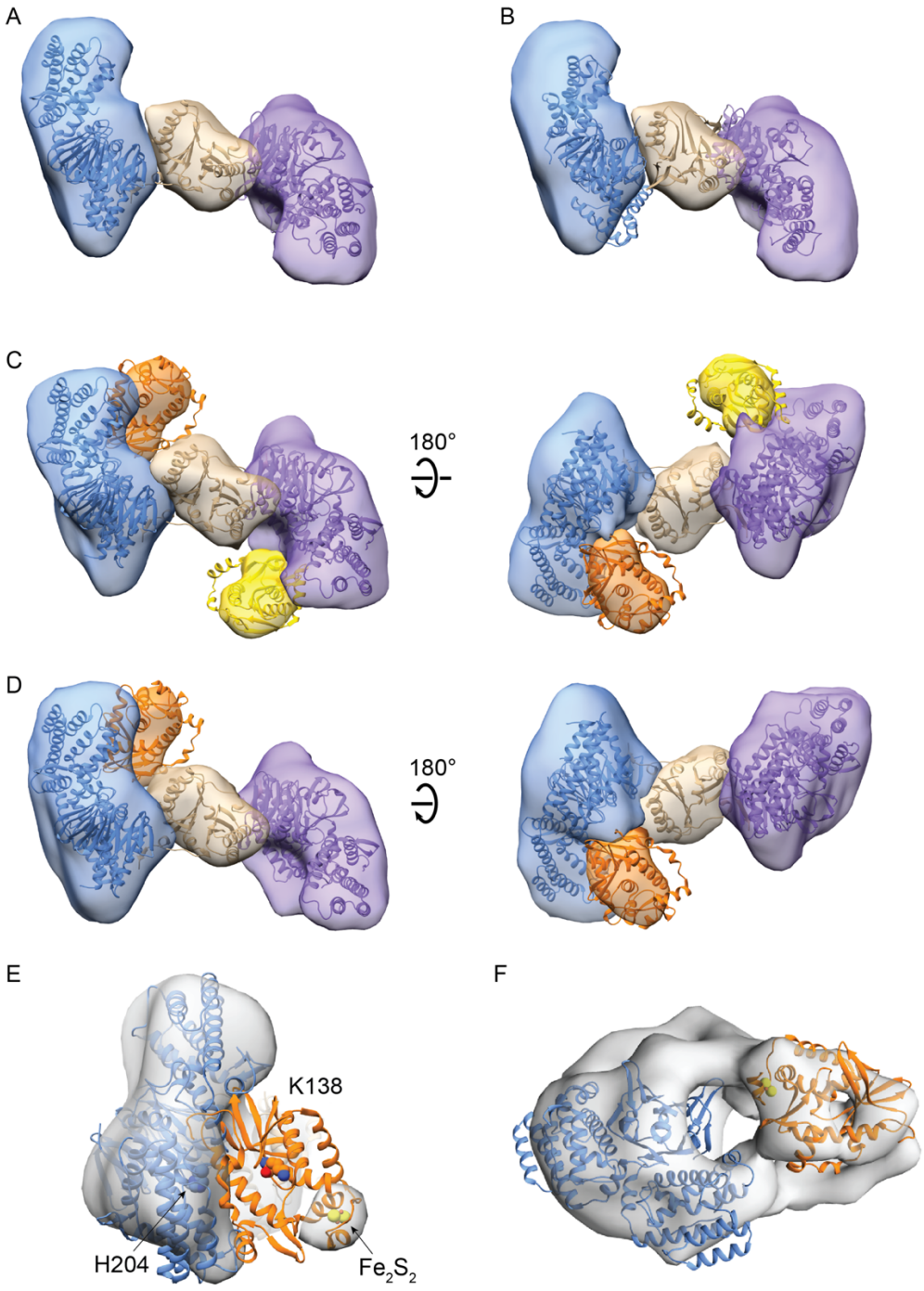

Figure 3. Architecture of Cas1-Cas2 and Cas4-Cas1-Cas2 complexes. (A) Segmented 3D reconstruction of Cas1-Cas2 with *E. faecalis* type II-A Cas1-Cas2 crystal structure docked (PDB: 5XVN). Cas1 dimers are shown in blue and purple and Cas2 dimer is shown in tan. (B) Segmented 3D reconstruction of Cas1-Cas2 with *E. coli* type I-E Cas1-Cas2 crystal structure docked (PDB: 5DS4). (C) Segmented 3D reconstruction of symmetrical Cas4-Cas1-Cas2 with *E. faecalis* Cas1-Cas2 and two copies of *P. calidifontis* Cas4 crystal structure (PDB: 4R5Q) docked. Cas4 is shown in orange and gold. (D) Segmented 3D reconstruction ofa symmetrical Cas4-Cas1-Cas2 with *E. faecalis* Cas1-Cas2 and one copy of *P. calidifontis* Cas4 crystal structure docked. (E) Close up of Cas1 and Cas4 with density of symmetrical Cas4-Cas1-Cas2 shown at lower contour level. The Fe_2S_2 cluster in the Cas4 structure was docked into the remaining Cas4 density observed at this contour level. The active site residues (Cas1 H204 and Cas4 K138) are shown as spheres. (F) 3D reconstruction of Cas4-Cas1 showing interface between the docked Cas1 dimer from PDB:5XVN and Cas4.

However, we observed similar amounts of processed prespacers with and without the CRISPR (Fig. 6B). Interestingly, when we tested cleavage of a single strand of the prespacer substrate, we observed a small amount of the cleavage product in the presence of the CRISPR DNA, while the DNA remained uncleaved in the absence of the CRISPR DNA or Cas4 (Fig. 6A, C). These data suggest that Cas4-Cas1-Cas2 can cleave ssDNA when a dsDNA is provided in *trans,* and that the duplex present within the prespacer is sufficient to stimulate complex formation.

To test these possibilities, we performed a cleavage assay with 5′ -end-^{32}P-labeled 25-nt ssDNA, while titrating a blunt-end 25-bp duplex similar to the dsDNA region of the prespacer substrate (Fig. 6D). Interestingly, at higher dsDNA concentrations, an increasing amount of the cleavage products was observed, whereas no detectable cleavage was observed without dsDNA (Fig. 6E). We also observed low levels of integration of ssDNA substrates into a mini-CRISPR target (Fig. 7). However, two integrated products were observed with or without Cas4, suggesting that the integration events with ssDNA may not be specific. Overall, these data show that ssDNA can be processed by the adaptation complex in the presence of any activator dsDNA and suggest that ssDNA could be used as prespacer substrates during adaptation.

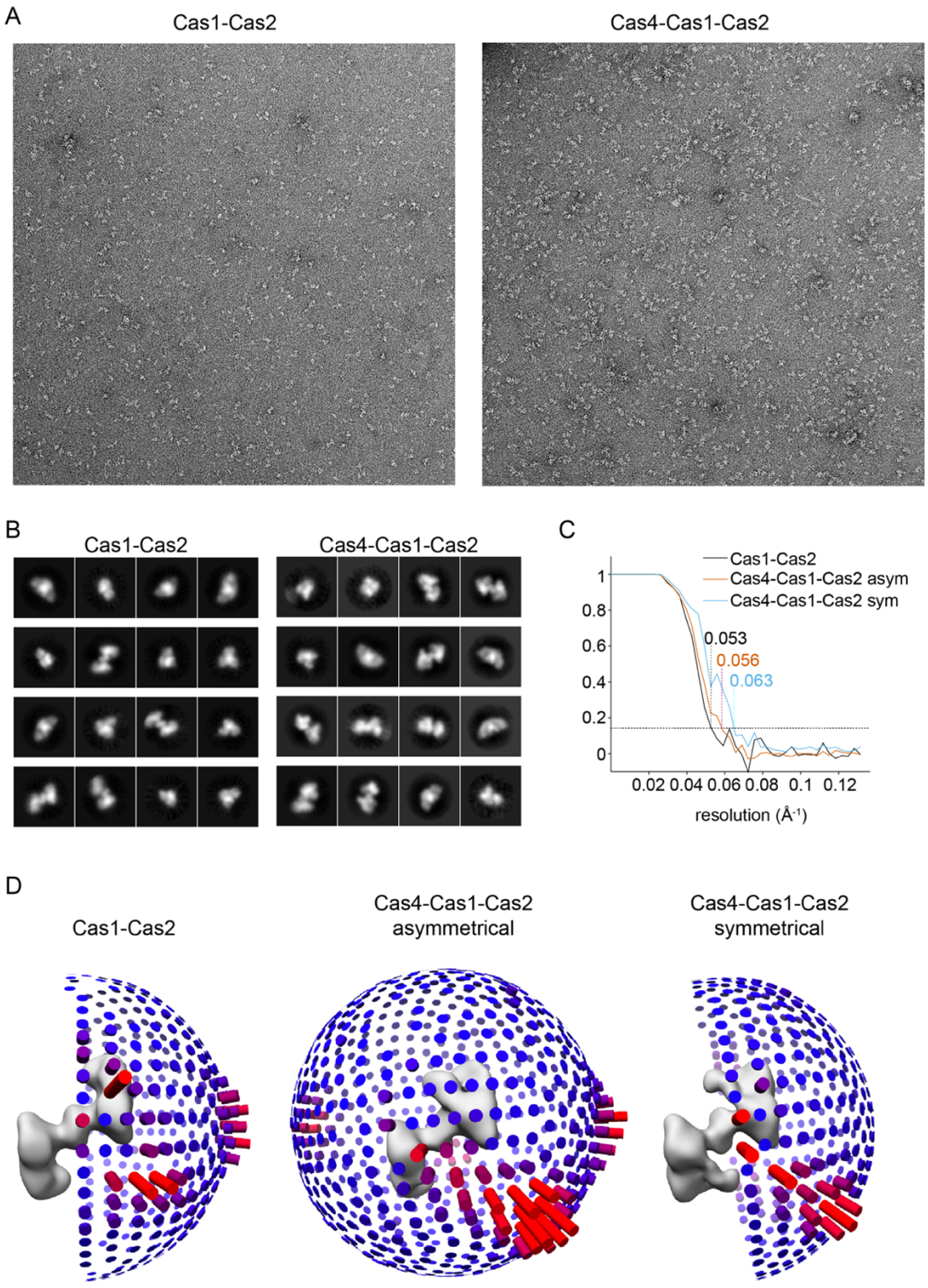

Figure 4. Single particle EM analysis of Cas1-Cas2 and Cas4-Cas1-Cas2. (A) Representative raw micrographs of Cas1-Cas2 and Cas4-Cas1-Cas samples. (B) Representative two-dimensional class averages of Cas1-Cas2 and Cas4-Cas1-Cas2. (C) Fourier shell correlation curves for Cas1-Cas2 (black), asymmetrical Cas4-Cas1-Cas2 (red) and symmetrical Cas4-Cas1-Cas2 (blue). The resolution using the gold-standard cutoff of 0.143 is indicated for each structure. (D) Angular distributions for the final reconstruction for each complex. Cas1-Cas2 and symmetrical Cas4-Cas1-Cas2 were refined with C2 symmetry.

Precise PAM-specific ssDNA processing by Cas4-Cas1-Cas2

We previously showed that Cas4 processes prespacers in a PAM-dependent manner (Lee et al., 2018), but the exact cleavage sites and specificity of Cas4 remain unclear. The observation that Cas4-Cas1-Cas2 can process ssDNA allowed us to more precisely define the cleavage site. We conducted prespacer processing assays using ssDNA substrates containing a 5′ -GAA-3′ PAM between T-rich sequences in the presence of activating dsDNA. Comparison with ddNTP Sanger sequencing reactions revealed that Cas4-Cas1-Cas2 precisely cleaved the ssDNA directly upstream of the PAM, while Cas1-Cas2 or Cas4 alone failed to cleave the substrates (Fig. 8A). This cleavage site is consistent with the expected processing site relative to the PAM required to form a functional spacer during spacer acquisition.

We next wondered whether the location of PAM sites affects the processing activity. We designed ssDNA substrates containing three PAM sites at varying intervals. Substrates with 10, 8, 6, 4 or 2-nt between three PAM sites generated three different sized cleavage products, indicating that the adaptation complex cleaved directly upstream of each PAM site (Fig. 9A-D). However, when we introduced three PAM sites consecutively on ssDNA substrates, we observed a predominant product at the first PAM position (Fig. 8B), indicating that processing was inhibited at the second and third PAM site. Together, these results suggest that PAM sites must be more than 2-nt apart for optimal processing.

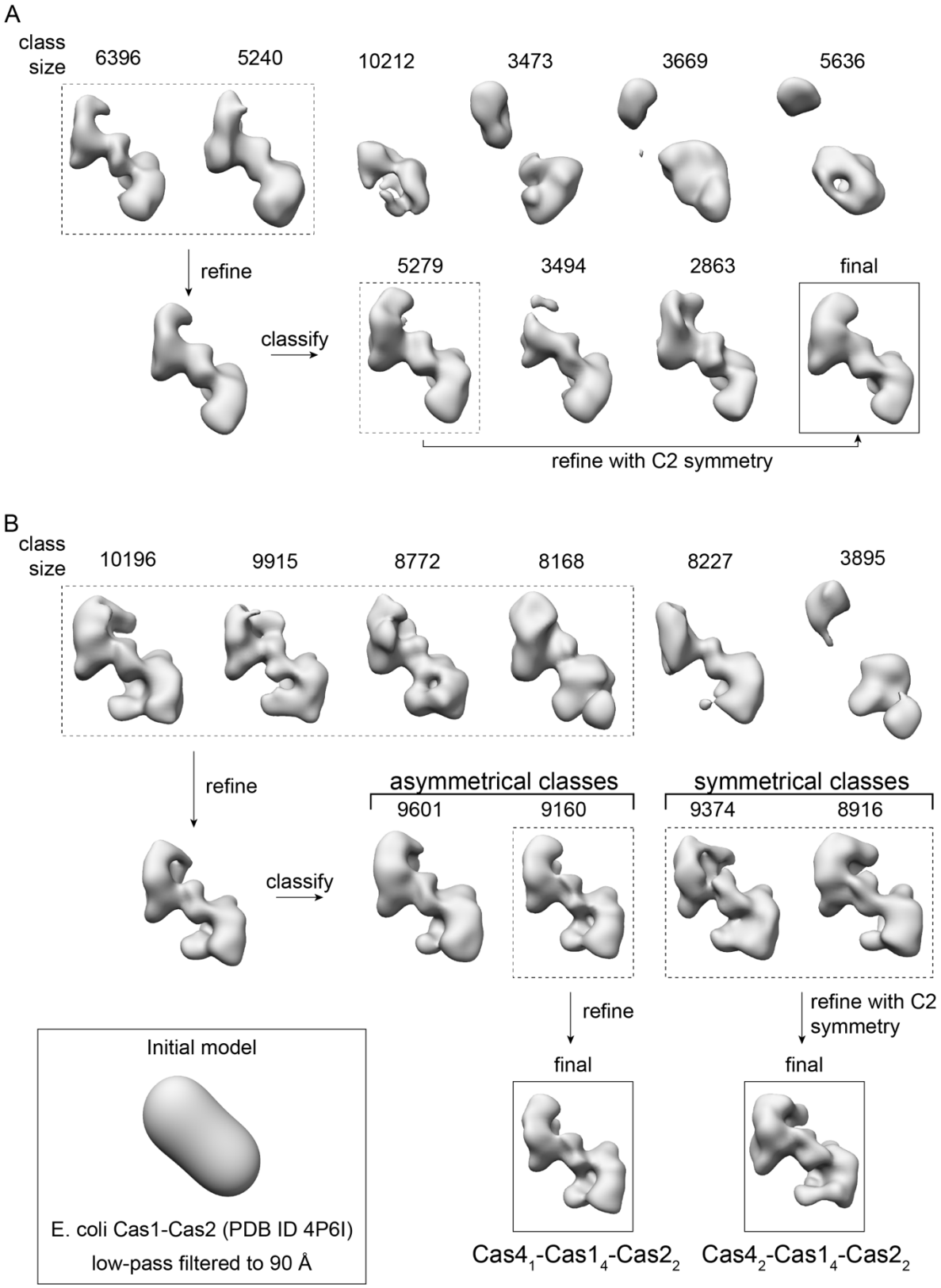

Figure 5. Three-dimensional classification of Cas1-Cas2 and Cas4-Cas1-Cas2. The starting model is shown in the lower left corner. The number of particles in each class is shown above the reconstruction.

To explore how the PAM-flanking regions affect cleavage, we used ssDNA substrates containing non-T-rich sequences on either side of the PAM (Fig. 10A). Notably, all substrates were cleaved directly upstream of the PAM site, indicating that Cas4 is highly PAM specific (Fig. 10B-D). Overall, these data suggest that the correct PAM sequences on ssDNA regions are required for processing activity by the adaptation complex.

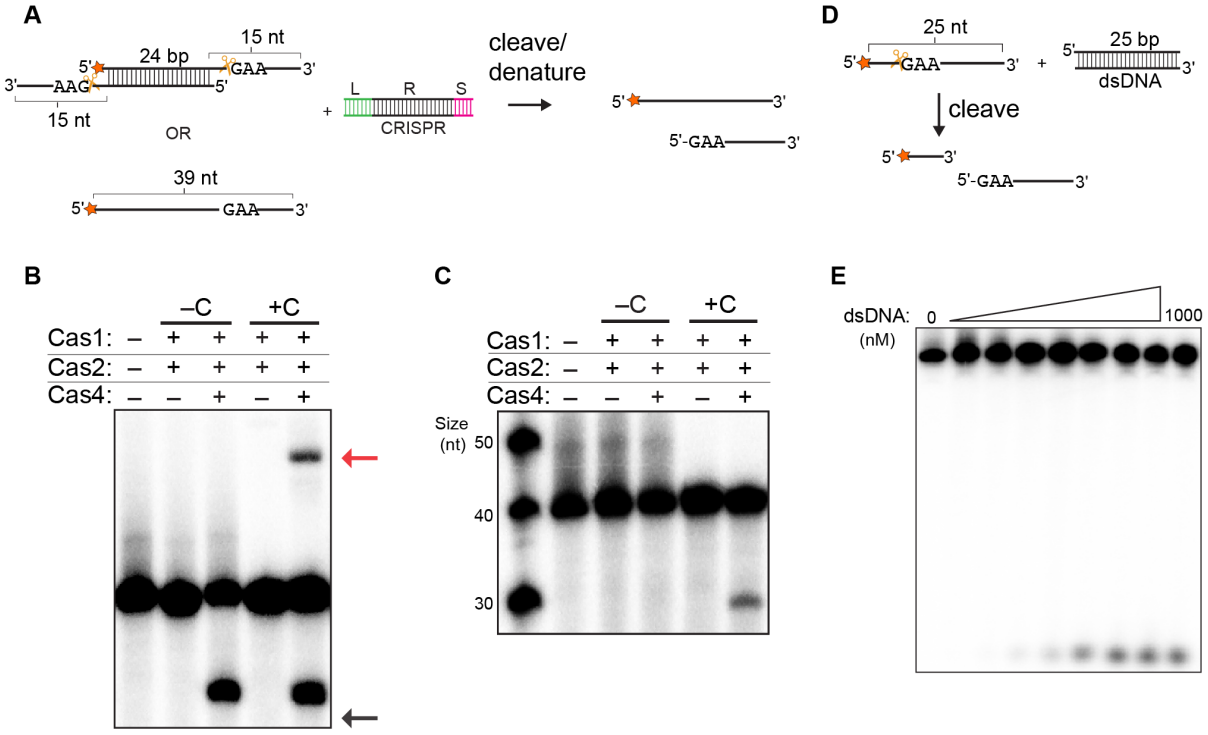

Figure 6. ssDNA processing by Cas4-Cas1-Cas2 complex. (A) Schematic view of prespacer cleavage assay for (B) and (C). L indicates leader, R indicates repeat, S indicates spacer in the CRISPR DNA substrate. (B) Prespacer processing assay using 24-bp duplex with 15-nt 3′ overhang in the absence or presence of CRISPR DNA labeld as C. Red arrow indicates integrated products, and black arrow indicates the processed prespacer. (C) Prespacer processing assay using 49-nt ssDNA in the absence or presence of CRISPR DNA. (D) Schematic view of prespacer assay using 25-bp duplex and 25-nt ssDNA. (E) Prespacer processing assay using 25-nt ssDNA with titration of 25-bp duplex. Radiolabel is indicated with a star. The dsDNA is titrated to 0, 1, 5, 10, 20, 50, 100, 500, and 1000 nM.

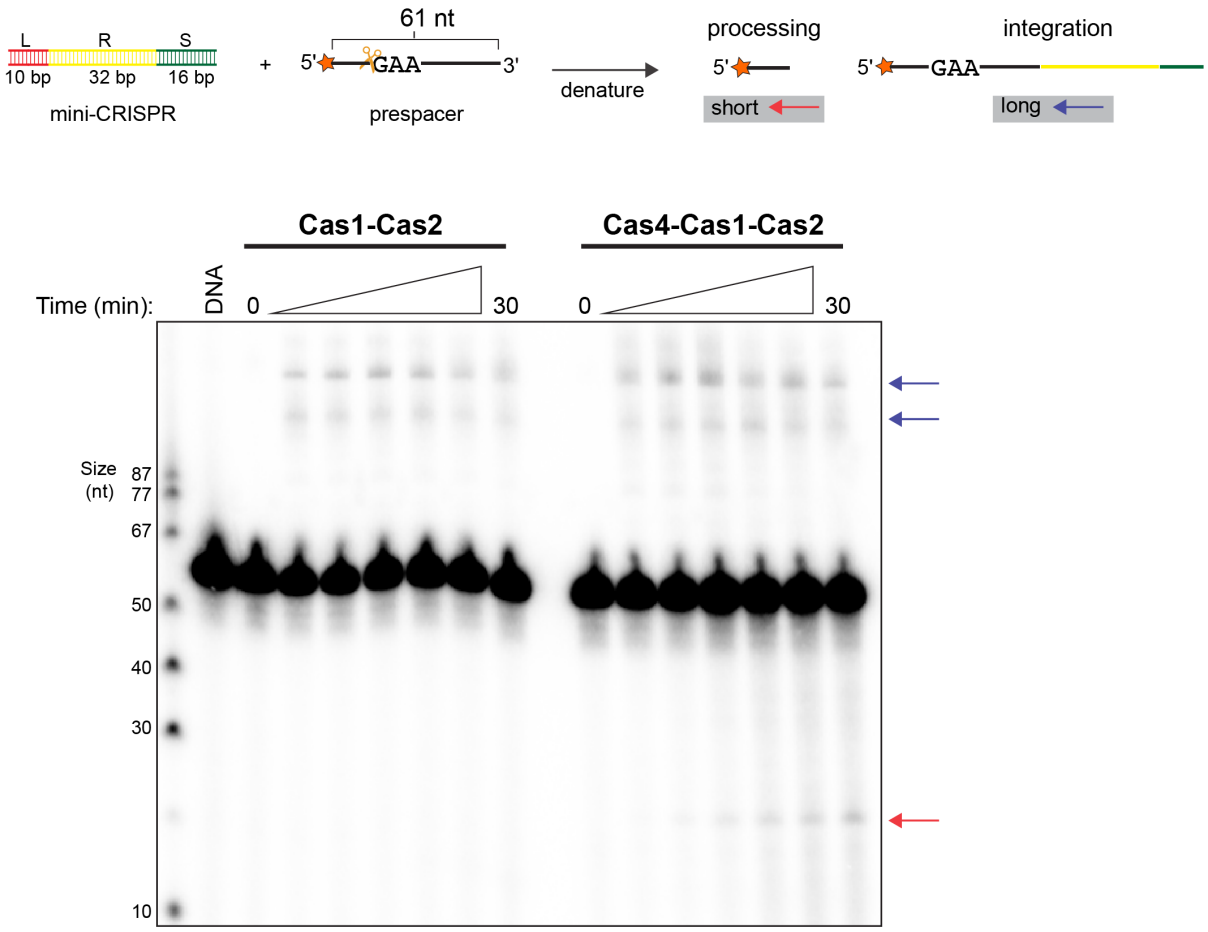

Figure 7. Cas4-Cas1-Cas2 complex integrates ssDNA into CRISPR locus. Integration using mini-CRISPR (2 μM) with 5′-radiolabeled ssDNA with or without Cas4 in Cas1-Cas2. Time points were taken at 0, 1, 2, 5, 10, 15, 30 min. Red arrow indicates the short products of processed prespacers. Blue arrows indicate the long products from integration events of unprocessed prespacers.

Discussion

Cas4 is a core family of CRISPR adaptation proteins, but its exact role and mechanism in spacer acquisition is relatively poorly understood. In particular, although there has been some preliminary biochemical evidence that Cas4 directly associates with Cas1-Cas2 to form a higher-order complex, these complexes were either very weak (Lee et al., 2018) or formed only under renaturing conditions (Plagens et al., 2012). Here, we discovered that the presence of dsDNA substrates stabilizes the formation of both Cas1-Cas2

and Cas4-Cas1-Cas2 complexes in the *B. halodurans* type I-C system. For the first time, we present the architectures of type I-C Cas1-Cas2 and Cas4-Cas1-Cas2 complexes that mediate prespacer selection, processing, and integration during CRISPR adaptation.

Our structural analysis of *B. halodurans* type I-C adaptation complexes reveal a structure that is mutually exclusive with the previously determined Cas4-Cas1 complex (Lee et al., 2018). In the Cas4-Cas1 complex, two Cas1 dimers are in close proximity and would exclude the Cas2 dimer. In addition, the interaction surface between Cas1 and Cas4 appears different in the two complexes. In Cas4-Cas1, the two Cas4 molecules each interact with one wing tip of the butterfly-like Cas1 dimers. In the Cas4-Cas1-Cas2 complex, Cas4 appears to interact along the length of one Cas1 wing. Modeling of crystal structures into the Cas4-Cas1-Cas2 reconstruction suggests that Cas4 and Cas1 interact extensively and that their active sites may be in close proximity. These structural results explain how Cas4 may hand off single-stranded ends of prespacer substrates following processing, as was previously proposed.

Previously, we showed that Cas4 processes prespacers with duplexes flanked by 3′ overhangs, and that processing activity was independent of the duplex or overhang lengths (Lee et al., 2018). Our current results reveal that the duplex region of these prespacers likely activates Cas4-Cas1-Cas2 for cleavage, and that a similar processing activity can be activated when the duplex and ssDNA are provided in trans. Processing is highly PAM-specific and occurs precisely upstream of the PAM, consistent with the expected processing position to form a functional prespacer. In these experiments, Cas4 did not appear to have strong cleavage activity at non-PAM sites, suggesting that Cas4-Cas1-Cas2 only processes the PAM-proximal end in type I-C. In type I-A from *Pyrococcus furiousus*, two distinct Cas4 proteins coordinate the processing of each end of the prespacer (Shiimori et al., 2018).

However, most Cas4-containing systems, including type I-C, lack a second *cas4* gene. It is possible that in these systems, Cas4 may define the PAM-distal end of the prespacer through an alternative cleavage activity or that another host factor is required for this processing activity.

Formation of functional spacers requires that the PAM-end of the prespacer must be integrated at the leader-distal end of the repeat following prespacer processing. It remains unclear how spacer orientation is defined during integration. In type I-E, one nucleotide of the PAM is retained following prespacer processing, and this nucleotide may help to define the PAM end during integration (Datsenko et al., 2012; Swarts et al., 2012). In the type I-A system, the two Cas4 proteins are required to define orientation, suggesting that their presence in an adaptation complex may define the orientation of the spacer. Notably, we observed asymmetrical complexes of Cas4-Cas1-Cas2 containing only one Cas4 subunit. This configuration, along with the hypothesis that Cas4 only processes the PAM-end of the prespacer, could suggest that only a single copy of Cas4 is required to form a functional adaptation complex. Indeed, the asymmetrical complex may also define prespacer orientation, based on which end of the prespacer is bound at the Cas4-end of the complex. Future studies will be required to determine the significance of asymmetrical Cas4-Cas1-Cas2 complexes for prespacer processing and the orientation of integration.

Interestingly, the type I-C Cas1-Cas2 complex integrated ssDNA efficiently but non-specifically into the CRISPR locus. Single-stranded substrates are not optimal for acquisition due to the lack of a second 3′ -OH nucleophile. Type I-E *E. coli* Cas1-Cas2 failed to integrate ssDNA, while type I-F Cas1-Cas2/3 had very low ssDNA integration (Nuñez et al., 2015a; Fagerlund et al., 2017). However, ssDNA were efficiently integrated in type III-B system with reverse transcriptase-Cas1 fusion protein and Cas2 (Silas et al., 2016). Together

with ssDNA processing by Cas4-Cas1-Cas2, this suggests that ssDNA could be used as a prespacer substrate during acquisition. It remains unclear how the second strand would be generated to form the full-site integration product.

Overall, our data supports a model for how the Cas4-Cas1-Cas2 complex mediates in prespacer processing and integration. The Cas4-Cas1 complex dissociates and reforms with Cas2 to form a higher-order complex in the presence of dsDNA. The Cas4-Cas1-Cas2 complex processes the single-stranded region of duplexes or ssDNA directly upstream of PAM sties and integrates into the CRISPR locus. Together, our findings reveal overall architectures of *B. halodurans* type I-C Cas4-Cas1-Cas2 and suggests similar structural constraints in other Cas4-containing systems.

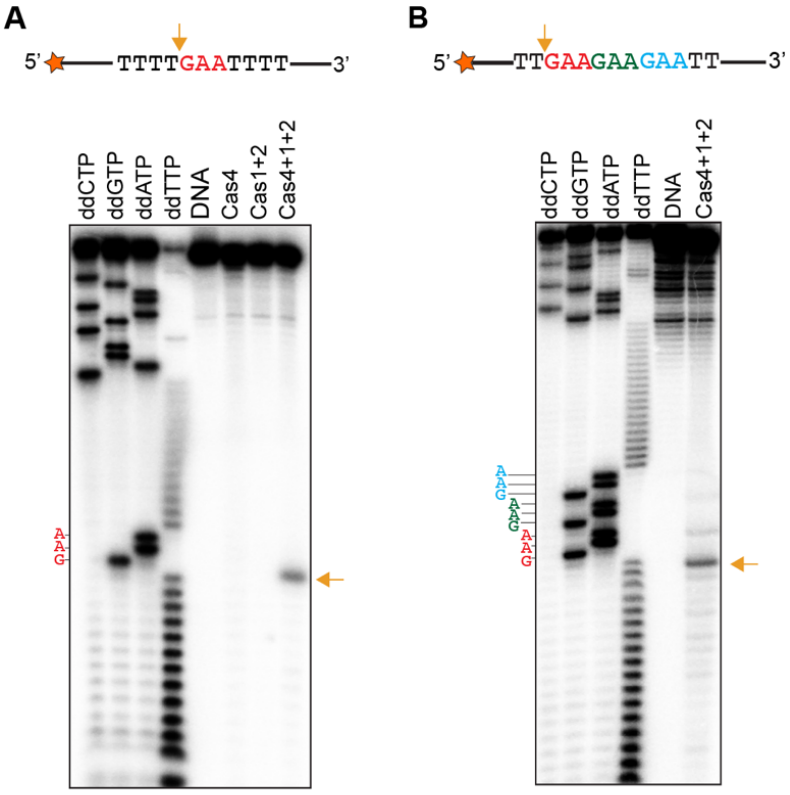

Figure 8. Cas4-Cas1-Cas2 processes directly upstream of the PAM site.
(A) Prespacer processing assay with ssDNA containing one PAM (GAA) site within multiple T sequences. (B) Prespacer processing assay with ssDNA containing three PAM sites consecutively on ssDNA substrates. First PAM site is labeled red, second PAM site is labeled green and the last PAM site is labeled cyan. Yellow arrows are indicated the predominant cleavage site.

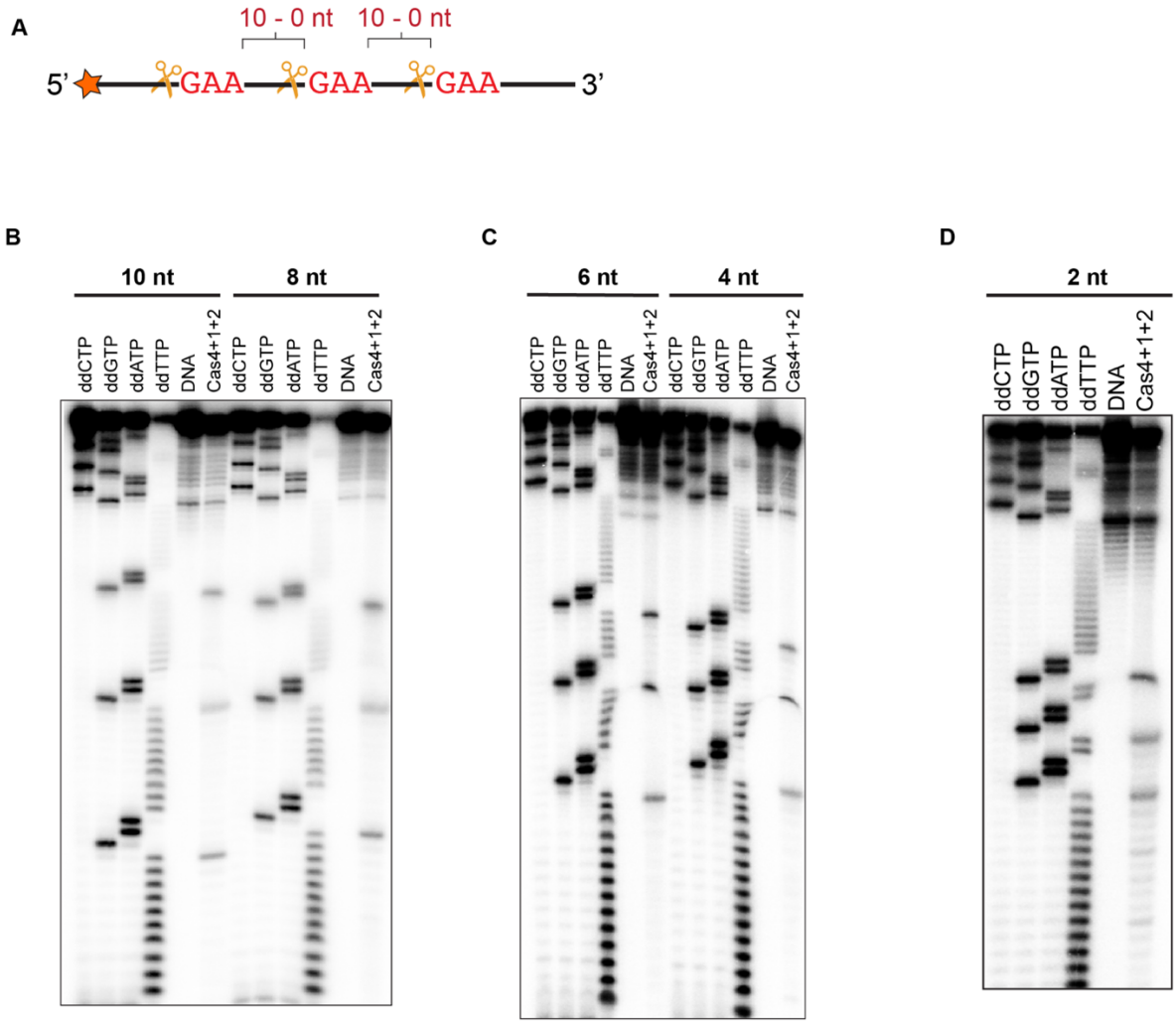

Figure 9. (A) Schematic view of ssDNA with three PAM sites with 10, 8, 6, 4, 2, or 0-nt length between the sites. (B) Prespacer processing assay with ssDNA substrates containing three PAM (GAA) sites that have 10-nt (left) or 8-nt (right) between the sites. (B) Prespacer processing assay with ssDNA substrates containing three PAM (GAA) sites with 6-nt (left) or 4-nt (right) between the sites. (C) Prespacer processing assay with ssDNA substrates containing three PAM (GAA) sites that are spaced in 2-nt between the sites.

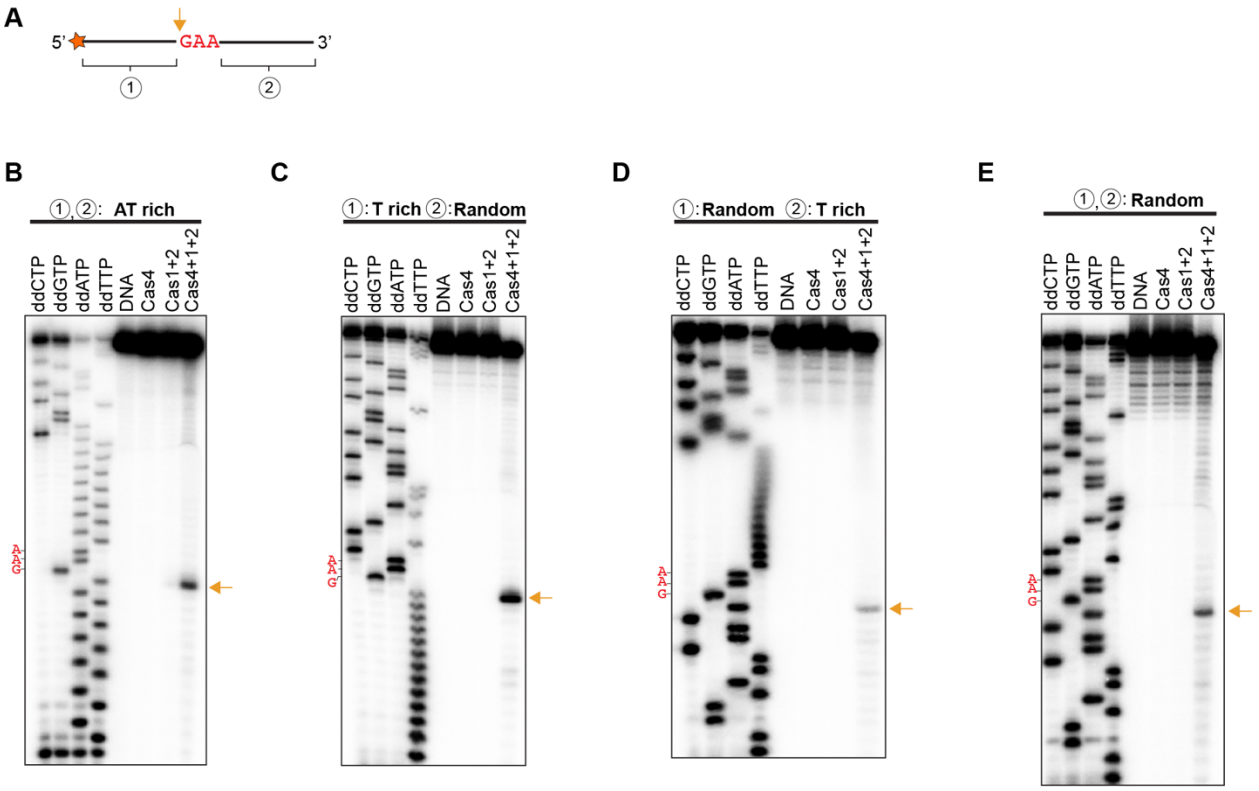

Figure 10. Cleavage of ssDNA substrates with different PAM-flanking regions. (A) Schematic view of ssDNA substrates containing one PAM (GAA) site with different sequences upstream (1) or downstream (2) of the PAM site. (B) Sequences upstream (1) and downstream (2) are AT-rich. (C) Sequences upstream (1) is T-rich and downstream (2) is random. (D) Sequences upstream (1) is random and downstream (2) is T-rich. (E) Sequences upstream (1) and downstream (2) are random.

Materials and Methods

Protein purification

Cas1, Cas2, and Cas4 were cloned from *Bacillus halodurans* and purified as previously described (Lee et al., 2018). For complex formation, Cas1, Cas2, and hairpin DNA substrates or Cas4, Cas1, Cas2, and hairpin DNA substrates were incubated in equal molar ratios (1:1:1 or 1:1:1:1) in dialysis buffer (20 mM HEPES (pH 7.5), 100 mM NaCl, 5 % glycerol, 2 mM DTT, and 2 mM $MnCl_2$) overnight at 4°C. The free Cas2 and free DNA were separated from the complex on a 5 mL HiTrap Q column (GE Healthcare). Fraction

containing all two or three proteins were pooled, concentrated, and further purified using a Superdex 75 10/30 GL column (GE Healthcare) in size exclusion buffer (20 mM HEPES (pH 7.5), 100 mM KCl, 5% glycerol, 2mM DTT, and 2 mM $MnCl_2$) for Cas1-Cas2-target complex and size exclusion buffer B (20 mM HEPES (pH7.5), 250 mM KCl, 5% glycerol, 2mM DTT, and 2mM $MnCl_2$) for Cas4-Cas1-Cas2-target complex.

DNA substrate preparation

All oligonucleotides were synthesized by Integrated DNA Technologies. Sequences of all DNA substrates are shown in Table 1. All DNA substrates were purified on 10% urea-PAGE. Double-stranded DNA was hybridized by heating to 95°C for 5 minutes followed by slow cooling to room temperature in oligo annealing buffer (20 mM HEPES (pH 7.5), 25 mM KCl, 10 mM $MgCl_2$). Prespacers were labeled with [γ-^{32}P]-ATP (PerkinElmer) and T4 polynucleotide kinase (NEB) for 5′-end labelling. Radiolabeled strands were purified using illustra microspin G-25 columns (GE healthcare).

Negative stain EM sample preparation and data collection

To prepare grids for EM imaging, Cas1-Cas2-target or Cas4-Cas1-Cas2-target was diluted to ~100 nM based on the calculations of proteins and hairpin target DNA concentrations, and 3 μL of sample was applied to a glow-discharged copper 400-mesh continuous carbon grid for one minute at room temperature. The excess sample was blotted with Whatman filter paper, followed by immediate application of 3 μL 2% (w/v) uranyl formate. The excess stain was blotted, followed by immediate application of 3 μL 2% uranyl formate. This step was repeated once more. The grids were allowed to dry for at least 5 minutes prior to imaging.

Images were collected on a 200 keV JEOL 2100 transmission electron microscope equipped with a Gatan OneView camera at a nominal magnification of 60,000x and pixel size of 1.9 Å. The electron dose was between 30-40 electrons/Å^2. For each sample, 200 images were collected manually at a defocus range of 1-2.5 µm.

Image processing and single-particle analysis

All image processing and analysis was performed in Scipion v. 1.2 (de la Rosa-Trevín et al., 2016) (available at http://scipion.i2pc.es/). The contrast transfer function (CTF) for each micrograph was estimated using CTFFIND4 (Rohou and Grigorieff, 2015). For each sample, ~200 particles were picked using Xmipp manual picking, followed by automated picking using the manually picked particles as a training set (Abrishami et al., 2013; Sorzano et al., 2013; Vargas et al., 2013). In total, 95,669 Cas1-Cas2 and 115,445 Cas4-Cas1-Cas2 particles were present in the initial data set. Particles were extracted using a 160 x 160 pixel box. To reduce computational requirements, the particles were down sampled by a factor of 2 to a final box size of 80 x 80 pixels (~152 x 152 Å). The particles were normalized and subjected to reference-free 2D classification using Relion 2.1 (Scheres, 2012) (Fig. 4B). The initial 100 class averages were inspected, and averages with clear density and representing different projections of the complex were selected for further analysis. These particles were subjected to a second round of 2D classification into 50 classes using Relion to further clean the particles. After selection of the best 2D classes, the Cas1-Cas2 dataset contained 34,626 particles, while the Cas4-Cas1-Cas2 dataset contained 49,173 particles.

Particles were next subjected to 3D classification in Relion using the X-ray crystal structure of *E. coli* Cas1-Cas2 (PDB: 4P6I (Nuñez et al., 2014)) low-pass-filtered to 90 Å as a starting model (Fig. 5). Each dataset was initially classified into 6 classes. For Cas1-Cas2, 11,636

particles from two similar 3D classes with the clearest density were combined (Fig. 5A). These particles were subjected to 3D refinement using Relion, and the refined volume was used to create a 3D mask. The refined particles were subjected to a second round of classification into three classes using the 3D mask as a reference mask (Fig. 5A). A class containing 5,279 particles with the clearest density and most symmetrical features was selected. These particles were subjected to 3D refinement while enforcing C2 symmetry and using the 3D mask.

For Cas4-Cas1-Cas2, 37,051 particles from four out of six initial 3D classes that appeared to contain Cas1-Cas2 with additional density were selected for further refinement (Fig. 5B). These particles were subjected to 3D refinement and then further classified into four 3D classes. The resulting classes had clearly defined extra density in comparison to the Cas1-Cas2 3D reconstruction. For two of these classes, the extra density was symmetrical and present extending from each Cas1 lobe, while for the other two classes, the extra density was only observed extending from one Cas1 lobe. Particles (18,290) from the two symmetrical 3D classes were combined and subjected to 3D refinement while enforcing C2 symmetry. Particles (9160) from one of the asymmetrical 3D classes were subjected to 3D refinement with C1 symmetry. The refined 3D reconstructions were used to create 3D masks, and each set of particles was subjected to a final round of refinement using the 3D mask as reference mask.

The resolutions of the final 3D reconstruction were 19.0 Å, 16.0 Å and 17.9 Å for Cas1-Cas2, symmetrical Cas4-Cas1-Cas2 and asymmetrical Cas4-Cas1-Cas2, respectively, based on Gold Standard Fourier Shell Correlation (FSC) at a cutoff of 0.143 (Fig. 4C). The Euler angle plots for the final 3D reconstructions revealed some preferred orientations but indicated a wide angular distribution in the data (Fig. 4D). Volumes were segmented using

Segger (Pintilie et al., 2010) in UCSF Chimera (Pettersen et al., 2004). For Cas1-Cas2, the protein subunits of the X-ray crystal structure of *E. faecalis* Cas1-Cas2 bound to prespacer (PDB: 5XVN (Xiao et al., 2017)) were docked into the final 3D reconstruction using Chimera Fit in Map function. For Cas4-Cas1-Cas2, the *E. faecalis* Cas1-Cas2 and *Pyrobaculum calidifontis* Cas4 (PDB: 4R5Q (Lemak et al., 2014)) were docked into the volumes.

Prespacer processing and integration assays

Prespacer processing assays were performed using 5′ -radiolabeled prespacer with 500 nM Cas4, 200 nM Cas1, 200 nM Cas2 in integration buffer (20 mM HEPES (pH 7.5), 100mM KCl, 5% glycerol, 2mM DTT, and 2 mM $MnCl_2$). The complexes were incubated on ice 10 min and all reactions were performed at 65°C for 20 min. Reactions were quenched with 2X RNA dye (NEB) and heat at 95°C for 5 min followed by ice cooling for 3 min. Samples were run on 12% urea-PAGE. The gels were dried and imaged using phosphor screens.

For sequencing reaction, Sequenase Version 2.0 DNA sequencing kit (Applied Biosystems) was used. Samples were prepared by hybridizing template with 5′-radiolabeled primer at 65 °C for 2 min and slowly cooling to RT within 30 minutes. Samples were incubated with the chain terminators (ddGTP, ddCTP, ddATP, and ddTTP) at 37°C for 15 minutes and quenched with 2X RNA dye. The cleaved products were prepared in the presence of 1.2 μM Cas4, 600 nM Cas1, 600 nM Cas2, and 1 μM of dsDNA substrates and quenched with 2X RNA dye. All samples were heat at 95°C for 5 min followed by ice cooling for 3 min prior to loading and run on 0.4 mm 8% urea-PAGE. The gels were dried and imaged using phosphor screens.

For integration assays, 2 μM mini-CRISPR array was used with 5′-radiolabeled single stranded DNA substrates with or without 1.5 μM Cas4 in the presence of 1 μM Cas1 and 1 μM Cas2 in integration buffer at 65°C for 30 min. Samples were quenched with 2X RNA dye and run on 6% urea-PAGE. The gels were dried and imaged using phosphor screens.

Table 1. Oligos used in this study. Bold indicates repeat sequences and italic indicates leader sequences. RC indicates the complementary strand of the previous strand.

Sequence (5′ → 3′)	Description
GCGTAGCTGAGGACCACCAGAACAGTTTTGAATTTTTTTTTTTTTTTTTT	25-nt 3′ overhang prespacer
CTGTTCTGGTGGTCCTCAGCTACGC	RC
GCGTAGCTGAGGACCACCAGAACAG	25-bp duplex
CTGTTCTGGTGGTCCTCAGCTACGC	RC
TTTTGAATTTTTTTTTTTTTTTTT	25-nt ssDNA
*GATTTTCGCT***GTCGCACTCTTCATGGGTGCGTGG ATTGAAAT**ATTGAGGTAGGTATTG	Mini-CRISPR array
CAATACCTACCTCAATATTTCAATCCACGCACCC ATGAAGAGTGCGACAGCGAAAATC	RC
GCGTAGCTGAGGACCTTTTTTTTTTTTTTGA ACTCGTATTCAACAGCAGGTCGACAAGCTTG	61-nt ssDNA prespacer
*GATTTTCGCT***GTCGCACTCTTCATGGGTGCGTGGATTGAAAT** ATTGAcgatagTCAATATTTCAATCCACGCACCCATGAAGAGTGC GACAGCGAAAATC	CRISPR hairpin target
CTAGTATGATCATGTCCAACGAATCAATACCTACCTCAATGAACGGAT	48-bp duplex
ATCCGTTCATTGAGGTAGGTATTGATTCGTTGGACATGATCATACTAG	RC
GCGTAGCTGAGGACCTTTTTTTTTTTTTTGAATTTTTTTTTTTTTTCAGGT CGACAAGCTTG	T-rich ssDNA prespacer
CAAGCTTGTCGACCTGAAAAAAAAAAAAAATTCAAAAAAAAAAAAA GGTCCTCAGCTACGC	RC
GCGTAGCTGAGGACCTTTTTTTTTTTTTTTTTTTTGAAGAAGAATTTTT TTTTTTTTTTTTTTGACAAGCTTGCGACA	3 PAM sites without spacing
TGTCGCAAGCTTGTCAAAAAAAAAAAAAAAAAAAATTCTTCTTCAAA AAAAAAAAAAAAAAAAAGGTCCTCAGCTACGC	RC
GCGTAGCTGAGGACCTTTTTTTTTTTGAATTTTTTTTTTGAATTTT TTTTTTGAATTTTTTTTTTGACAAGCTTGCGACA	3 PAM sites interspersed in 10-nt
TGTCGCAAGCTTGTCAAAAAAAAAATTCAAAAAAAAAATTCAAAAA AAAAATTCAAAAAAAAAAGGTCCTCAGCTACGC	RC
GCGTAGCTGAGGACCTTTTTTTTTTTTTGAATTTTTTTTGAATTTTTTTT GAATTTTTTTTTTTTGACAAGCTTGCGACA	3 PAM sites interspersed in 8-nt
TGTCGCAAGCTTGTCAAAAAAAAAAAATTCAAAAAAAATTCAAAAAAAA TTCAAAAAAAAAAAAGGTCCTCAGCTACGC	RC
GCGTAGCTGAGGACCTTTTTTTTTTTTTTTGAATTTTTTGAATTTT TTGAATTTTTTTTTTTTTTGACAAGCTTGCGACA	3 PAM sites interspersed in 6-nt
TGTCGCAAGCTTGTCAAAAAAAAAAAAAATTCAAAAAATTCAAAA AATTCAAAAAAAAAAAAAGGTCCTCAGCTACGC	RC

Table 1. (continued)

GCGTAGCTGAGGACCTTTTTTTTTTTTTTTTGAATTTTGA ATTTTGAATTTTTTTTTTTTTTTTGACAAGCTTGCGACA	3 PAM sites interspersed in 4-nt
TGTCGCAAGCTTGTCAAAAAAAAAAAAAAAAATTCAAAATTCAAAA TTCAAAAAAAAAAAAAAAAGGTCCTCAGCTACGC	RC
GCGTAGCTGAGGACCTTTTTTTTTTTTTTTTTTGAATTGAATTGA ATTTTTTTTTTTTTTTTTTGACAAGCTTGCGACA	3 PAM sites interspersed in 2-nt
TGTCGCAAGCTTGTCAAAAAAAAAAAAAAAAAATTCAATTCAATTCA AAAAAAAAAAAAAAAAAGGTCCTCAGCTACGC	RC
GCGTAGCTGAGGACCTATATATATATATGAATATATATATATATA CAGGTCGACAAGCTTG	AT-rich ssDNA prespacer
CAAGCTTGTCGACCTGTATATATATATATATTCATATATATAT ATAGGTCCTCAGCTACGC	RC
GCGTAGCTGAGGACC TTGGT ATTCA ACAGA ATTTT TTTTT TTTTTCAGGT CGACA AGCTT G	Non-T-rich on upstream / T-rich on downstream ssDNA prespacer
CAAGCTTGTCGACCTGAAAAAAAAAAAAAATTCTGTTGAATACCAAG GTCCTCAGCTACGC	RC
GCGTAGCTGAGGACC TTTTT TTTTT TTTGA ACTCG TATTC AACAG CAGGT CGACA AGCTT G	T-rich on upstream / non-T-rich downstream ssDNA prespacer
CAAGCTTGTCGACCTGCTGTTGAATACGAGTTCAAAAAAAAAAAAA GGTCCTCAGCTACGC	RC
GCGTAGCTGAGGACC TTGGT ATTCA ACAGA ACTCG TATTC AACAGCAGGT CGACA AGCTT G	Non-T-rich on up- and downstream ssDNA prespacer
CAAGCTTGTCGACCTGCTGTTGAATACGAGTTCTGTTGAATACCAA GGTCCTCAGCTACGC	RC
GCGTAGCTGAGGACC	Primer used for ddNTP Sanger sequencing

References

Abrishami, V. et al. (2013). A pattern matching approach to the automatic selection of particles from low-contrast electron micrographs. Bioinformatics *29*, 2460–2468.

Barrangou, R. et al. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. Science (80-.). *315*, 1709–1712.

Bolotin, A. et al. (2005). Clustered regularly interspaced short palindrome repeats (CRISPRs) have spacers of extrachromosomal origin. Microbiology *151*, 2551–2561.

Brouns, S.J.J. et al. (2008). Small CRISPR RNAs Guide Antiviral Defense in Prokaryotes Stan. Science (80-.). *321*, 960–964.

Charpentier, E. et al. (2015). Biogenesis pathways of RNA guides in archaeal and bacterial CRISPR-Cas adaptive immunity. FEMS Microbiol. Rev. 1–14.

Datsenko, K.A. et al. (2012). Molecular memory of prior infections activates the CRISPR / Cas adaptive bacterial immunity system. Nat. Commun. *3*, 945–947.

Deveau, H. et al. (2008). Phage response to CRISPR-encoded resistance in Streptococcus thermophilus. J. Bacteriol. *190*, 1390–1400.

Fagerlund, R.D. et al. (2017). Spacer capture and integration by a type I-F Cas1–Cas2-3 CRISPR adaptation complex. Proc. Natl. Acad. Sci. 201618421.

Hochstrasser, M.L., and Doudna, J.A. (2014). Cutting it close: CRISPR-associated endoribonuclease structure and function. Trends Biochem. Sci. *40*, 58–66.

Hudaiberdiev, S. et al. (2017). Phylogenomics of Cas4 family nucleases. Evol. Biol. *17*, 232.

Ivančić-Baće, I. et al. (2015). Different genome stability proteins underpin primed and naïve adaptation in E. coli CRISPR-Cas immunity. Nucleic Acids Res. *43*, 10821–10830.

Jackson, S.A. et al. (2017). CRISPR-Cas: Adapting to change. Science (80-.). *356*, eaal5056.

Kieper, S.N. et al. (2018). Cas4 Facilitates PAM-Compatible Spacer Selection during CRISPR Adaptation. Cell Rep. *22*, 3377–3384.

Koonin, E. V. et al. (2017). Diversity, classification and evolution of CRISPR-Cas systems. Curr. Opin. Microbiol. *37*, 67–78.

de la Rosa-Trevín, J.M. et al. (2016). Scipion: A software framework toward integration, reproducibility and validation in 3D electron microscopy. J. Struct. Biol. *195*, 93–99.

Lee, H. et al. (2018). Cas4-Dependent Prespacer Processing Ensures High-Fidelity Programming of CRISPR Arrays. Mol. Cell 1–12.

Lemak, S. et al. (2014). The CRISPR-associated Cas4 protein Pcal 0546 from Pyrobaculum calidifontis contains a [2Fe-2S] cluster : crystal structure and nuclease activity. Nucleic Acids Res. *42*, 11144–11155.

Li, M. et al. (2014). Adaptation of the Haloarcula hispanica CRISPR-Cas system to a purified virus strictly requires a priming process. Nucleic Acids Res. *42*, 2483–2492.

Marraffini, L.A. (2015). CRISPR-Cas immunity in prokaryotes. Nature *526*, 55–61.

Marraffini, L.A., and Sontheimer, E.J. (2008). CRISPR Interference Limits Horizontal Targeting DNA. Science (80-.). *322*, 1843–1845.

Mojica, F.J.M. et al. (2005). Intervening sequences of regularly spaced prokaryotic repeats derive from foreign genetic elements. J. Mol. Evol. *60*, 174–182.

Nogales, E., and Scheres, S.H.W. (2015). Cryo-EM: A Unique Tool for the Visualization of Macromolecular Complexity. Mol. Cell *58*, 677–689.

Nuñez, J.K. et al. (2014). Cas1-Cas2 complex formation mediates spacer acquisition during CRISPR-Cas adaptive immunity. Nat. Struct. Mol. Biol. *21*, 528–534.

Nuñez, J.K. et al. (2015a). Integrase-mediated spacer acquisition during CRISPR–Cas adaptive immunity. Nature.

Nuñez, J.K. et al. (2015b). Foreign DNA capture during CRISPR–Cas adaptive immunity. Nature *527*, 535–538.

Pettersen, E.F. et al. (2004). UCSF Chimera - A visualization system for exploratory research and analysis. J. Comput. Chem. *25*, 1605–1612.

Pintilie, G.D. et al. (2010). Quantitative analysis of cryo-EM density map segmentation by watershed and scale-space filtering, and fitting of structures by alignment to regions. J. Struct. Biol. *170*, 427–438.

Plagens, A. et al. (2012). Characterization of the CRISPR/Cas subtype I-A system of the hyperthermophilic crenarchaeon Thermoproteus tenax. J. Bacteriol. *194*, 2491–2500.

Pourcel, C. et al. (2005). CRISPR elements in Yersinia pestis acquire new repeats by preferential uptake of bacteriophage DNA, and provide additional tools for evolutionary studies. Microbiology *151*, 653–663.

Redding, S. et al. (2015). Surveillance and Processing of Foreign DNA by the Escherichia coli CRISPR-Cas System Article Surveillance and Processing of Foreign DNA by the Escherichia coli CRISPR-Cas System. Cell 1–12.

Rohou, A., and Grigorieff, N. (2015). CTFFIND4: Fast and accurate defocus estimation from electron micrographs. J. Struct. Biol. *192*, 216–221.

Rollie, C. et al. (2015). Intrinsic sequence specificity of the Cas1 integrase directs new spacer acquisition. Elife *4*.

Rollie, C. et al. (2017). Prespacer processing and specific integration in a type I-A CRISPR system. Nucleic Acids Res. *46*, 1007–1020.

Sashital, D.G. et al. (2012). Mechanism of Foreign DNA Selection in a Bacterial Adaptive Immune System. Mol. Cell *46*, 606–615.

Scheres, S.H.W. (2012). RELION: Implementation of a Bayesian approach to cryo-EM structure determination. J. Struct. Biol. *180*, 519–530.

Semenova, E. et al. (2011). Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. Proc. Natl. Acad. Sci. *108*

Shiimori, M. et al. (2018). Cas4 Nucleases Define the PAM, Length, and Orientation of DNA Fragments Integrated at CRISPR Loci. Mol. Cell *70*, 814–824.e6.

Silas, S. et al. (2016). Direct CRISPR spacer acquisition from RNA by a natural reverse transcriptase-Cas1 fusion protein. Science (80-.). *351*.

Sorzano, C.O. et al. (2013). Semiautomatic, High-Throughput, High-Resolution Protocol for Three-Dimensional Reconstruction of Single Particles in Electron Microscopy. Humana Press, 171–193.

Sternberg, S.H. et al. (2014). DNA interrogation by the CRISPR RNA-guided endonuclease Cas9. Nature *507*, 62–67.

Swarts, D.C. et al. (2012). CRISPR interference directs strand specific spacer acquisition. PLoS One *7*, 1–7.

Vargas, J. et al. (2013). Particle quality assessment and sorting for automatic and semiautomatic particle-picking techniques. J. Struct. Biol. *183*, 342–353.

Wang, J. et al. (2015). Structural and Mechanistic Basis of PAM-Dependent Spacer Acquisition in CRISPR-Cas Systems. Cell *163*, 840–853.

Xiao, Y. et al. (2017). How type II CRISPR–Cas establish immunity through Cas1–Cas2-mediated spacer integration. Nature 1–23.

CHAPTER 4. DNA TARGETING BY TYPE I-C CASCADE AND CONSTRUCTION OF MINIMAL-CYSTEINE VARIANT FOR SMFRET

Introduction

Prokaryotes use adaptive immune systems comprising clustered regularly interspaced short palindromic repeats (CRISPR) arrays and CRISPR-associated (*cas*) genes to defend against infection (Barrangou et al., 2007; Brouns et al., 2008). CRISPR-Cas immunity occurs in three stages. First, the Cas1-Cas2 integrase captures and inserts short fragments from foreign nucleic acids as spacers in the CRISPR array (Yosef et al., 2012; Nuñez et al., 2015a, 2015b). Then, the CRISPR array is transcribed and processed into short CRISPR RNA (crRNAs), each containing a single spacer flanked by partial repeat sequences (Hochstrasser and Doudna, 2015). Next, the crRNAs assemble with Cas proteins to form a crRNA-guided surveillance complex to recognize and destroy targets bearing complementarity to the crRNA (Marraffini and Sontheimer, 2008; Wiedenheft et al., 2011a; Westra et al., 2012).

CRISPR-Cas systems are diverse and can be classified into two classes, six types, and many subtypes based on the architecture and composition of *cas* gene loci (Mohanraju et al., 2016; Koonin et al., 2017). Type I systems are the most abundant in nature, found in both cultivated and uncultivated genomes of bacteria and archaea (Makarova et al., 2015; Burstein et al., 2016). Type I systems encode multiple proteins to form a crRNA-mediated interference complex that targets DNA (Jackson and Wiedenheft, 2015; Koonin et al., 2017). Recent structures from several subtypes have revealed that the type I systems share similarities in overall architecture, suggesting a common origin (Jackson et al., 2014; Mulepati et al., 2014; Zhao et al., 2014; Gao et al., 2016; Hayes et al., 2016; Hochstrasser et al., 2016; Chowdhury et al., 2017; Xiao et al., 2017, 2018).

In the type I-E system found in *E. coli*, Cascade is a 400-kDa surveillance complex and includes one crRNA and five Cas proteins. Cas5e and Cas6e cap each end of the crRNA

and six copies of Cas7 oligomerize along the crRNA backbone. Cas8e and two copies of Cas11 interact with Cas5e and Cas7, respectively (Jackson et al., 2014; Mulepati et al., 2014; Zhao et al., 2014). In order to recognize the target, Cascade searches for PAM (protospacer adjacent motif) sequences that can be found adjacent to the target region (Mojica et al., 2009; Redding et al., 2015; Xue et al., 2017). The PAM is important for the complex to be able to bind the target, and is also required to distinguish foreign DNA from the host genome (Semenova et al., 2011). Single molecule studies show that *E. coli* Cascade samples PAM sequences rapidly through three-dimensional diffusion (Redding et al., 2015; Xue et al., 2017), while *T. fusca* Cascade scans via facilitated diffusion (Brown et al., 2017). When Cascade encounters a bona fide target flanked with a canonical PAM, Cascade unwinds the double stranded DNA and the crRNA base pairs with the target strand to form an R-loop structure. The formation of the complete R-loop triggers the recruitment of Cas3 nuclease helicase to degrade the target DNA (Hayes et al., 2016; Xiao et al., 2017, 2018).

Type I-E has been studied the most extensively due to the use of *E. coli* as a model organism, and is also the most abundant system found in sequenced bacteria. Although type I-C is the second most abundant system, it is relatively understudied. Type I-C system is unique in that it only requires three proteins to form a surveillance complex, rather than five proteins in type I-E system(Makarova et al., 2015; Koonin et al., 2017). Most type I systems encode a Cas6 endoribonuclease for processing pre-crRNAs and binding to a stem loop region as a part of the Cascade complex (Gesner et al., 2011; Sashital et al., 2011; Jackson et al., 2014; Mulepati et al., 2014; Zhao et al., 2014). However, the type I-C system lacks Cas6 and uses Cas5c instead for cleaving pre-crRNAs, which is non-catalytic but has a similar structural role in the Cascade complexes from other subtypes (Nam et al., 2012). In type I-E Cascade, the Cas8e large subunit recognizes the PAM sequence while two copies of the

Cas11 small subunit stabilize the displaced non-target strands (Sashital et al., 2012; Hayes et al., 2016; Xue et al., 2016, 2017, Xiao et al., 2017, 2018). But, this separately encoded composition of the large and small subunit is only found in type I-A and I-E. Other subtypes such as type I-C instead utilize a fusion of Cas8e and Cas11 into a single protein encoded by the *cas8* gene (Makarova et al., 2015; Koonin et al., 2017).

A recent study of *D. vulgaris* type I-C system showed that three proteins are necessary and sufficient for forming a DNA-targeting surveillance complex with one copy of each Cas5c and Cas8c, and seven copies of Cas7 (Hochstrasser et al., 2016). When type I-C Cascade is bound to the target duplex, the similarities between the overall folds of Cas8c and Cas8e and orientation of target and non-target strand in type I-C and type I-E structures suggest that Cas8c recognizes the PAM sequences and stabilizes the R-loop structure (Hochstrasser et al., 2016). Despite the similarities between type I-C and type I-E, it remains unstudied how type I-C Cascade searches for target using a fusion of large and small subunit, Cas8c.

Here, we have investigated binding interactions of the type I-C Cascade from *Bacillus halodurans* with dsDNA. We observed that Cascade/I-C binds to the target DNA specifically at high pH, low salt, and high temperatures. Unlike *E. coli* Cascade/I-E, the Cascade/I-C complex shows stronger affinity for non-target strand or dsDNA targets containing non-canonical PAMs. These observations suggest that Cascade/I-C may use a target search mechanism involving longer-lived interactions with non-PAM sites. To test this, we have begun to develop a smFRET assay that enables direct visualization of how Cascade searches and binds to dsDNA. We were able to construct a cysteine-free version of preCascade (Cas5c-Cas7 complex) allowing for introduction of cysteine residues for site-specific fluorophore labeling. Using this system, we observed bulk FRET between Cy3-labelled

dsDNA target and Cy5-labelled Cascade upon DNA binding, demonstrating that labeled Cascade/I-C retains DNA binding activity. This labeling strategy will be used for future experiments using smFRET to study the Cascade/I-C target search mechanism.

Results

Construction of type I-C Cascade for binding assay

B. halodurans contains a type I-C CRISPR system, in which type I-C specific *cas* genes can be found upstream of CRISPR locus 4 in the genome. We wondered whether Cas5c and Cas7 form a preCascade complex in the absence of Cas8c. A recent biochemical study of Cas8c showed that it has the lowest affinity to crRNAs or other components within the complex (Hochstrasser et al., 2016), suggesting that Cas8c may readily dissociate from Cascade, as previously observed for Cas8e from *E. coli* type I-E Cascade (Sashital et al., 2012). Cas8 subunits can be added in excess of the rest of the complex to ensure its presence in the complex and we wished to develop a similar system for *B. halodurans* Cascade/I-C. We cloned Cas5c and Cas7 into pET28b and pCDF, respectively and co-expressed with pRepeat (a plasmid with first repeat-spacer-repeat under control of a T7 promoter) to purify preCascade. Cas8c was cloned into pET52b and purified separately using HisPur Ni NTA affinity. We also cloned Cas5c, Cas8c, and Cas7 in a single operon into pET28b and co-expressed with pRepeat to purify the full Cascade complex (Fig. 1A).

We then used electrophoretic mobility shift assays (EMSA) to analyze the binding interactions of each complex with target DNA. Because *Bacillus halodruans* strains are facultive alkaliphiles and polyextremophiles, we investigated the effect of temperature, pH, and salt concentration on binding activity. We incubated Cascade with pTarget containing a

canonical PAM sequence with a protospacer complementary to the crRNA spacer region (Fig 1B). The canonical PAM sequences in type I-C has been characterized as 5′-GAA-3′ on the target strand (Leenay et al., 2016; Rao et al., 2016). We observed binding at low Cascade concentration for pH 7.5-9.0, low salt and higher temperatures (Fig. 1C). The observation of improved binding at higher temperature suggests that thermal energy may facilitate DNA unwinding by Cascade (Fig. 1B). Interestingly, we observed a higher mobility band caused by supershifting of the plasmid at 100 nM concentration. Similar supershifting has been observed with *E. coli* Cascade, although only at higher concentrations (1 μM) (Xue et al., 2015). These data suggest that Cascade/I-C has stronger affinity for non-specific sites in the dsDNA than Cascade/I-E, resulting in multiple copies of the complex binding at sites around the plasmid even at low complex concentration.

Cascade/I-C binds non-specific targets

To further investigate the binding interactions, we used 5′-end ^{32}P-labeled DNA for binding assays using EMSA. We first tested binding of Cascade to ssDNA corresponding either to the target strand that is complementary to the crRNA, or the nontarget strand. As seen in *E. coli* Cascade (Jore et al., 2011; Sashital et al., 2012; van Erp et al., 2015), Cascade has high binding affinity for single-stranded target DNA containing a sequence complementary to the spacer region of the crRNA and a canonical PAM (5′-GAA-3′) sequence. At lower Cascade concentration, we observed several slower migrating bands that could be the transition states from unbound to partially bound states or subcomplexes binding to the ssDNA. Substrates were fully bound at 50 nM Cascade. No detectable binding events were observed at low concentration with the non-target strand; however, Cascade bound to

single-stranded non-target DNA at 50 nM (Fig. 2A). These data suggest that Cascade has nonspecific interactions with ssDNA but binds to the target strand more specifically than the non-target strand.

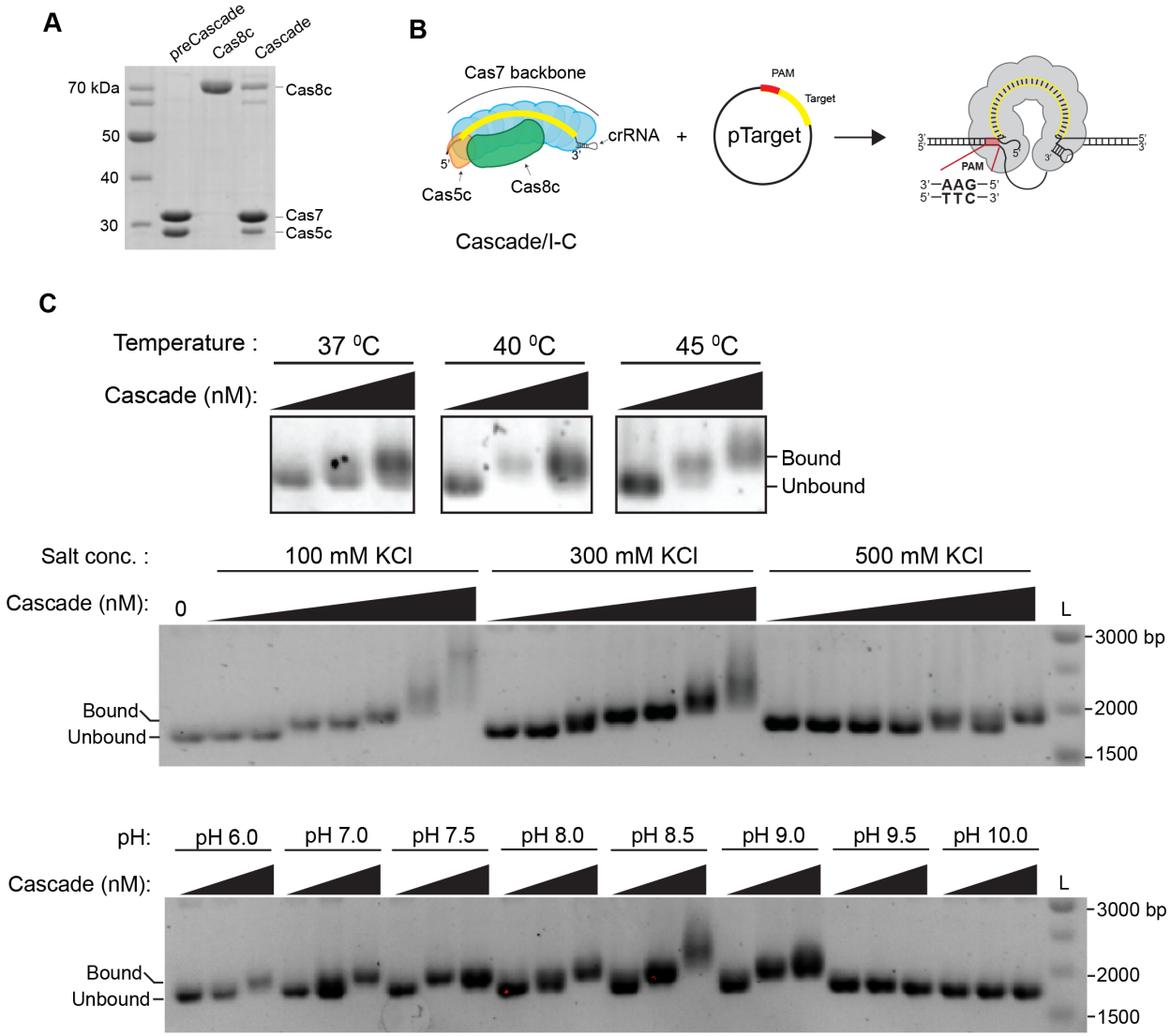

Figure 1. Cascade binds to target plasmid at low salt, neutral pH and high temperature. (A)Coomassie-blue stained SDS/PAGE gel of purified proteins. (B) Schematic view of plasmid binding assay. Cascade/I-C was incubated with pTarget containing canonical PAM (5′-GAA-3′) and target. (C) Plasmid binding assay by Cascade under different temperature, salt, and pH conditions. For temperature and pH conditions, 10, 50 and100nM Cascade were tested. For salt conditions 1, 10, 20, 40, 60, 100 and200 nM Cascade were tested.

We next tested binding activity using a 77-bp dsDNA target containing a canonical PAM. Cascade bound to the duplex at 50 nM, but the addition of saturating amounts of Cas8c increased the binding affinity, with nearly complete binding detected at 10 nM

Cascade concentration (Fig. 2B). These results suggest that an excess amount of Cas8c is necessary for dsDNA binding by Cascade because Cas8c readily dissociates at low concentrations of the complex and its dissociation limits dsDNA binding, similar to *E. coli* type I-E Cascade (Sashital et al., 2012).

The 'PAM scan' mechanism is predominant in CRISPR systems, in which the surveillance complex recognizes target sequences by first locating the adjacent PAM (Redding et al., 2015; Sternberg et al., 2015; Brown et al., 2017; Xue et al., 2017; Singh et al., 2018). Mutations in PAM sequences slow the searching process and attenuate binding affinity (Sashital et al., 2012; Rollins et al., 2015; van Erp et al., 2015; Hayes et al., 2016). We wondered whether mutations in the PAM cause binding defects in type I-C systems and tested binding affinity using substrates containing a reverse PAM (5′-TTC-3′) or a repeat complementary PAM (5′-AAT-3′) sequences. The last three nucleotides of the *B. halodurans* repeat are 5′-AAT-3′ and are predicted to protect the CRISPR array from self-targeting, as has been observed for the *E. coli* type I-E system (Sashital et al., 2012; Westra et al., 2013; Xue et al., 2015). Surprisingly, these PAM mutations caused only 2-10 folds reductions in binding activity compared to the perfect target (Fig. 2B-C). It is unclear whether PAM mutant binding is dictated by non-specific interactions, as observed with binding to the non-target strand, or through promiscuity for PAM recognition.

Construction of smFRET assay for Cascade-target binding

In contrast to Cascade/I-E, Cascade/I-C shows substantial non-specific binding interaction with dsDNA, suggesting that Cascade/I-C may interact more strongly with dsDNA during PAM scanning. To analyze PAM scanning in detail, we began development

of a single-molecule fluorescence resonance energy transfer (smFRET) assay to measure the association of Cascade/I-C with dsDNA (Fig. 3A). Single-molecule studies have been widely used to study dynamics of binding interactions between Cascade/I-E and dsDNA targets (Szczelkun et al., 2014; Blosser et al., 2015; Redding et al., 2015; Rutkauskas et al., 2015; Xue et al., 2016, 2017). We wished to develop a smFRET system to detect how individual Cascade complexes bind to dsDNA targets in real-time using Cy5 and Cy3 FRET pairs within Cascade and the dsDNA target, respectively (Fig. 3A). For the dsDNA, one strand of the DNA was synthesized containing a Cy3 dye at the 5′-end. In order to label the protein with Cy5, we designed a cysteine-free version of preCascade and Cas8c for site-specific labeling of either Cas5c or Cas8c. By introducing a cysteine on a surface-exposed site in the Cys-free background, Cy3- or Cy5- maleimide dyes can be incorporated site-specifically at the thiol group of the cysteine. Cys-free preCascade purified similar to WT, while Cys-free Cas8c aggregated during purification. We therefore only proceeded with the Cys-free preCascade construct.

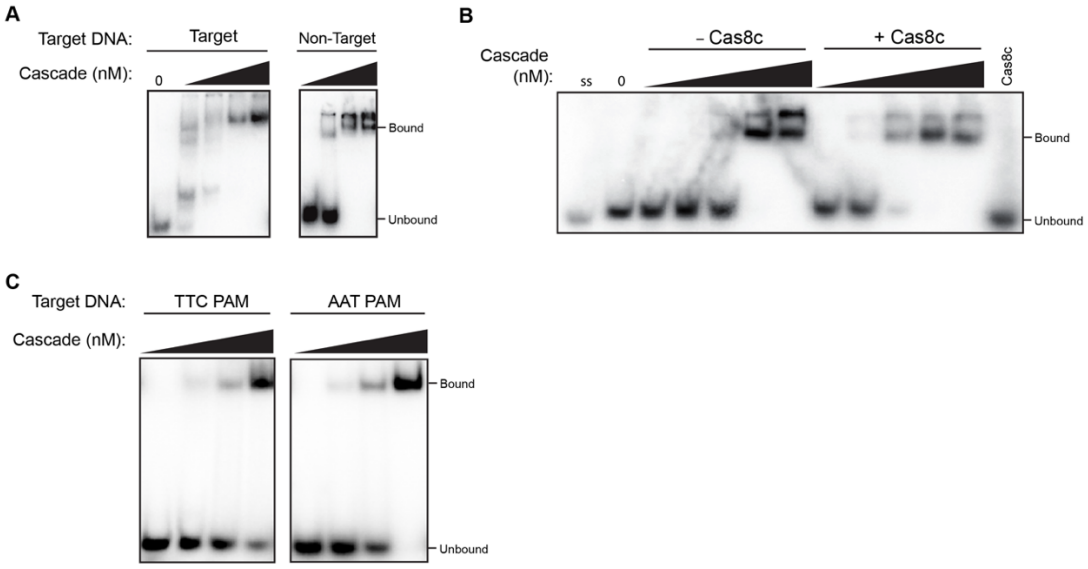

Figure 2. Gel-shift assay of Cascade. (A) Cascade (10, 20, 50, 100 nM) binding assay with single stranded target and nontarget strand. (B) Cascade (1, 10, 20, 50, 100 nM) binding assay with duplex target strand with the canonical PAM (GAA) and with excess Cas8c (1 μM). (C) Cascade (0.1, 10, 50, 100 nM) binding assay with TTC and AAT PAM duplex strand.

Energy transfer is dependent on the distances between donor and acceptor within ~10 nm. We selected Asp107 in Cas5c for mutation to Cys, as this site is surface exposed and predicted to be within 10 nm of the end of a dsDNA target based on the cryo-EM structure of Cascade/I-C and the crystal structure of *B. halodurans* Cas5c (Nam et al., 2012; Hochstrasser et al., 2016) (Fig. 3A). PreCascade with D107C Cas5c was successfully labeled with Cy5 (Fig. 3B). To test whether the placement of the dye pairs would result in FRET, we measured Cy5-labeled Cascade binding to Cy3-labeled DNA by EMSA. We observed that Cy5 labeled preCascade in the presence of Cas8c (1 uM) bound to Cy3 labeled dsDNA target strand based on Cy3-Cy5 FRET detected for the complex, indicating that the FRET pairs are in close proximity (Fig. 3C).

Discussion

Unlike the well-studied Cascade/I-E, Cascade/I-C is unique in that it only requires three proteins to form the surveillance complex. In this chapter, we began to investigate the similarities and differences between Cascade complexes from type I-C and I-E. We reconstituted a preCascade complex containing Cas5c, Cas7 subunits and crRNA and Cascade complex including Cas5c, Cas7, and Cas8c subunits. Interestingly, Cascade/I-C from *B. halodurans* requires high temperature for target binding that may help unwind the dsDNA (Fig. 1B-C). Addition of Cas8c in excess amount is also necessary to facilitate dsDNA binding because Cas8c readily dissociates from the complex at low concentration as previously observed for type I-E Cascade (Sashital et al., 2012; Hochstrasser et al., 2016) (Fig. 2B). Finally, we observed that at higher concentration, Cascade complexes were fully

bound to the substrate despite lack of complementarity to crRNAs or the presence of non-canonical PAMs (Fig. 2C). These results suggest that, unlike *E. coli* Cascade, Cascade/I-C may have strong non-specific interactions with DNA.

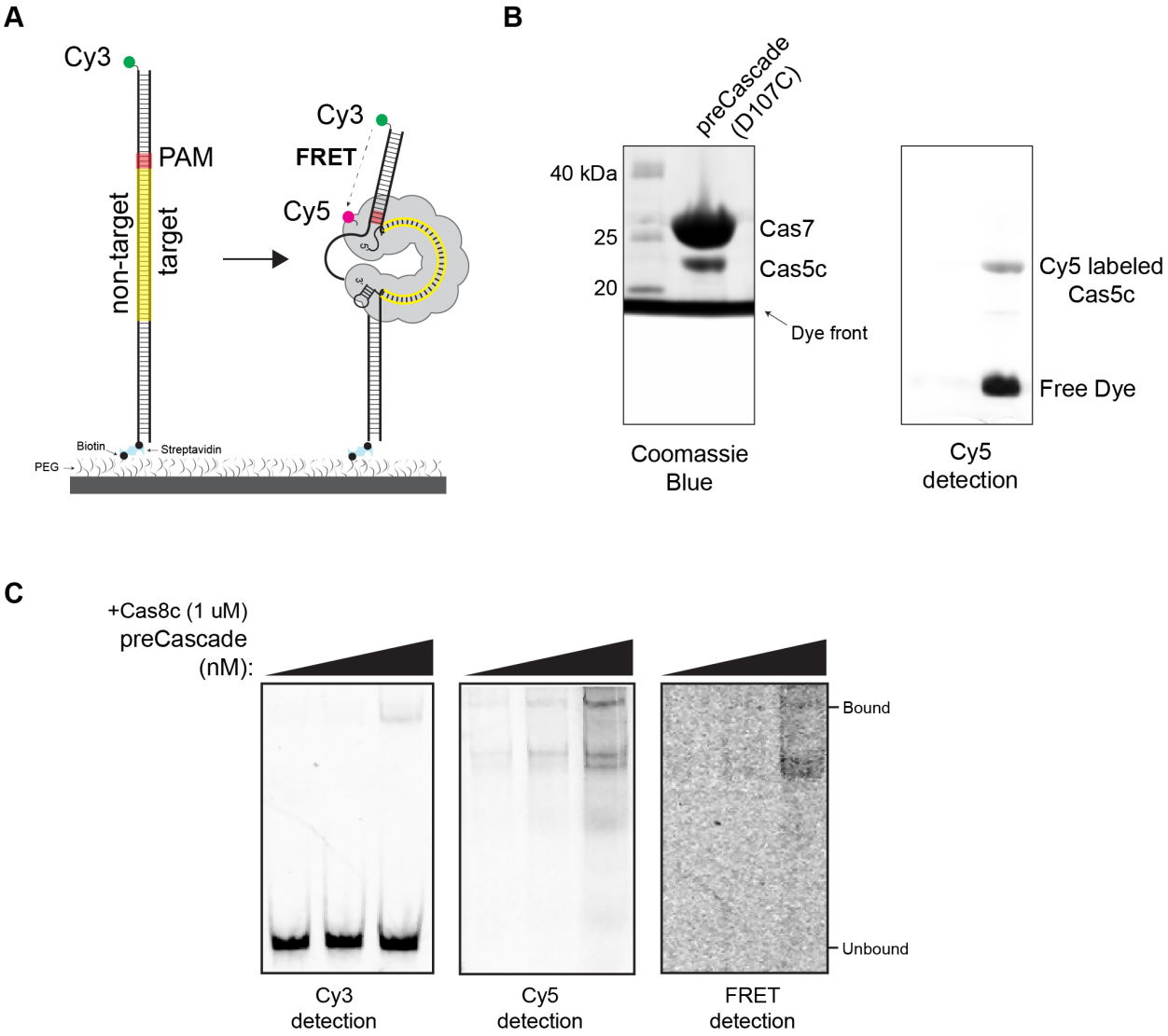

Figure 3. Construction of type I-C Cascade binding assay for smFRET. (A)Schematic view of smFRET experiments. (B) SDS PAGE gel with Cy5 labeling (left) and Coomassie Blue staining (Right). (C) Cy5 labeled preCascade (20, 50, 100 nM) binding assay with excess Cas8c (1 μM) with Cy3 labeled duplex target strand (100 nM).

Because the complex showed stronger affinity to non-specific DNA, we began to develop a construct for smFRET experiments to directly observe the dynamics of how individual Cascade complexes bind to dsDNA targets (Fig. 3A). Bulk labeling experiment

showed that the fluorophores were conjugated site-specifically on Cas5c and binding experiment using Cy3-labeled dsDNA and Cy5-labeled Cascade showed that the FRET pairs are in close proximity and ready for smFRET assay. We speculate that because of its stronger affinity to non-specific DNA, Cascade/I-C may dwell longer to search for PAMs on dsDNA, potentially through a sliding mechanism. Because *E. coli* Cascade searches for PAMs thru 3D diffusion via nonspecific interactions, this resulted in short-lived FRET events (Xue et al., 2017). Based on our observation of strong non-specific interactions, we predict that FRET events would be longer lived for Cascade/I-C. Future studies will use the smFRET system developed in this chapter to determine the kinetics and mechanisms of how Cascade/I-C binds and searches for dsDNA targets. In addition to providing fundamental information on how different surveillance complexes search for target DNA, this may provide insights to use this compact complex for biotechnology applications, such as gene repression (Leenay et al., 2016).

Materials and Methods

Cloning

Genomic DNA from *Bacillus halodurans* was obtained from ATCC. The *cas5c, cas7, and cas8c* genes were PCR amplified from the genomic DNA. For co-expression with N-terminal His_6-tagged *cas5c*, all three genes were PCR amplified as a single operon and cloned into pET28b. For individual expression, *cas5c* was cloned into pET28b, *cas7* into pCDF, and *cas8c* into pACYCDuet-1 and pET28b. pRepeat was generated by PCR amplification of the CRISPR array (first repeat-spacer-repeat array) from the genomic DNA

and ligated into BamHI- and EcoRI-digested pUC19 with T7 promoter and terminator. D107 in *cas5c* was mutated to cysteine.

Protein purification

His_6-tagged Cas8c was overexpressed in NiCo21(DE3) and grown to 0.6 OD_{600} in LB media, followed by overnight induction at 16°C with 0.5 mM IPTG. The cells were harvested, resuspended in Ni-NTA buffer (50 mM Na_2HPO_4 pH 8.0, 500 mM NaCl, 5% glycerol, 2 mM DTT) supplemented with 10 mM imidazole (pH 8.0) and 100 mM PMSF, and lysed using a homogenizer. All proteins were initially purified using HisPur Ni-NTA affinity resin using Ni-NTA buffer supplemented with 25 mM imidazole during washing or 250 mM imidazole during elution. Eluted samples were passed through a Chitin resin to remove non-specific proteins.

For Cascade, pET28b Cas5c-Cas7-Cas8c with pUC19 pRepeat or pET28b Cas5c, pCDF Cas7, pACYCDuet-1 Cas8c with pUC19 pRepeat were overexpressed in NiCo21(DE3) and grown to 0.7-0.8 OD_{600} in LB media, followed by overnight induction at 16°C with 1 mM IPTG. Cascade was purified using HisPur Ni-NTA affinity resin and Chitin resin as described above. For preCascade, pET28b Cas5c, pCDF Cas7 with pACYC pRepeat and pUC19 pRepeat were overexpressed and grown to 0.7-0.8 OD_{600} in LB media, followed by overnight induction at 16 °C with 1 mM IPTG. PreCascade was purified using HisPur Ni-NTA affinity resin and Chitin resin.

All proteins were further purified using a Superdex 200 10/30 in a size exclusion buffer (20 mM HEPES (pH 7.5), 100 mM KCl, 5% glycerol and 2 mM DTT).

Binding assay

For plasmid binding assays, the plasmid containing PAM sequences and protospacer, pTarget (200 ng (~ 5 nM)), was added to the Cascade complex at indicated concentration. Samples were incubated at 45°C for 20 min in a size exclusion buffer. Loading buffer (final concentration 50 mM EDTA (pH 8.0), 5% glycerol) was added to each sample. Samples were run on 0.7% unstained agarose gel at 18V overnight and post-stained with SYBRSafe for 30 min before visualization.

For oligo binding assays, Cascade alone or Cascade complex with 1 μM Cas8c were incubation on ice for 10 min in a size exclusion buffer. Oligos were labeled with [γ-^{32}P]-ATP and T4 polynucleotide kinase for 5′-end labeling. Oligos used are indicated in Table 1. 1 nM of labeled DNA was mixed with Cascade at indicated concentration and incubated at 45°C for 30 min. Loading buffer (final concentration 50 mM EDTA (pH 8.0), 5% glycerol) was added to each sample. Samples were run on 6% THE (20 mM Tris (pH 8.0), 20 mM HEPES (pH 8.0), 10mM EDTA) PAGE gel at 200V 2 hours at 4ºC. The gels were dried and imaged using phosphor screens on a Typhoon imager.

For Cy3-labeled oligo binding assay, 100 nM of Cy3-labeled oligo (IDT) was used with Cy5-labeled preCascade and wild-type Cas8c. Cy5-labeled preCascade was titrated with constant concentration of 1 μM Cas8c. Samples were incubated at 25°C for 30 min. Loading buffer (final concentration 50 mM EDTA (pH 8.0), 5% glycerol) was added to each sample and samples were run on 6% THE PAGE gel at 4ºC. The gels were visualized on a Typhoon imager using Cy3, Cy5 or Cy3-Cy5 FRET settings.

Cy5 labeling

D107C preCascade was labeled by Cy5-maleimide in the dark at 4 °C for 2 h in labelling buffer (20 mM HEPES (pH 7.5), 100 mM KCl, 5% glycerol, 5% DMSO and 1 mM TCEP). Free dye was removed from Cy5-labeled preCascade using a 3K protein concentrator. The labelling efficiency of preCascade was calculated by measuring the protein concentration at 280 nm and the dye concentration at 650 nm.

Table 1. Oligos used for binding assay

Name	Sequence (5′ → 3′)	Description
HL483	CATGAGGTCCTCGTCTAGTATGATCATGTCCAACGAATCAATACCTACCTCAAT GAACGGATGTACGATCAACACTC	Target with GAA PAM
HL484	GAGTGTTGATCGTACATCCGTTCATTGAGGTAGGTATTGATTCGTTGGACATGAT CAT ACTAGACGAGGACCTCATG	RC
HL548	CATGAGGTCCTCGTCTAGTATGATCATGTCCAACGAATCAATACCTACCTCAAT TTC CGGATGTACGATCAACACTC	Target with TTC PAM
HL549	GAGTGTTGATCGTACATCCGGAAATTGAGGTAGGTATTGATTCGTTGGACATGAT CATACTAGACGAGGACCTCATG	RC
HL550	CATGAGGTCCTCGTCTAGTATGATCATGTCCAACGAATCAATACCTACCTCAAT TAA CGGATGTACGATCAACACTC	Target with TAA PAM
HL551	GAGTGTTGATCGTACATCCGTTAATTGAGGTAGGTATTGATTCGTTGGACATGAT CATACTAGACGAGGACCTCATG	RC
HL866	ATACTGTGATTGGTAGACTGCG AA /3Bio/	Biotin-labeled strand
HL867	CGCAGTCTACCAATCACAGTATATCGTCTAGTATGATCATGTCCAACGAATCAAT ACCTACCTCAATGAACGGATGTACGATCAACACTG	Target strand
HL868	5'-cy3/CAGTGTTGATCGTACATCCGTTCATTGAGGTAGGTA TTGATTCGTTGGACATGATCAT ACTAGACGAT	Cy3 labeled on non-target strand

References

Barrangou, R. et al. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. Science (80-.). *315*, 1709–1712.

Blosser, T.R. et al. (2015). Two distinct DNA binding modes guide dual roles of a CRISPR-cas protein complex. Mol. Cell *58*, 60–70.

Brouns, S.J.J. et al. (2008). Small CRISPR RNAs Guide Antiviral Defense in Prokaryotes Stan. Science (80-.). *321*, 960–964.

Brown, M.W. et al. (2017). Assembly and translocation of a CRISPR-Cas primed acquisition complex. Cell 1–13.

Burstein, D. et al. (2016). Major bacterial lineages are essentially devoid of CRISPR-Cas viral defence systems. Nat. Commun. *7*, 1–8.

Chowdhury, S. et al. (2017). Structure Reveals Mechanisms of Viral Suppressors that Intercept a CRISPR RNA-Guided Surveillance Complex. Cell *169*, 47–57.e11.

Gao, P. et al. (2016). Type v CRISPR-Cas Cpf1 endonuclease employs a unique mechanism for crRNA-mediated target DNA recognition. Cell Res. *26*, 901–913.

Gesner, E.M. et al. (2011). Recognition and maturation of effector RNAs in a CRISPR interference pathway. Nat. Struct. Mol. Biol. *18*, 688–692.

Hayes, R.P. et al. (2016). Structural basis for promiscuous PAM recognition in type I–E Cascade from E. coli. Nature 1–16.

Hochstrasser, M.L. et al. (2016). DNA Targeting by a Minimal CRISPR RNA-Guided Cascade. Mol. Cell *63*, 840–851.

Hochstrasser, M.L., and Doudna, J.A. (2015). Cutting it close: CRISPR-associated endoribonuclease structure and function. Trends Biochem. Sci. *40*, 58–66.

Jackson, R.N. et al. (2014). Structural biology. Crystal structure of the CRISPR RNA-guided surveillance complex from Escherichia coli. Science (80-.). *345*, 1473–1479.

Jackson, R.N., and Wiedenheft, B. (2015). A Conserved Structural Chassis for Mounting Versatile CRISPR RNA-Guided Immune Responses. Mol. Cell *58*, 722–728.

Jore, M.M. et al. (2011). Structural basis for CRISPR RNA-guided DNA recognition by Cascade. Nat. Struct. Mol. Biol. *18*, 529–536.

Koonin, E. V. et al. (2017). Diversity, classification and evolution of CRISPR-Cas systems. Curr. Opin. Microbiol. *37*, 67–78.

Leenay, R.T. et al. (2016). Identifying and Visualizing Functional PAM Diversity across CRISPR-Cas Systems. Mol. Cell 1–11.

Makarova, K.S. et al. (2015). An updated evolutionary classification of CRISPR–Cas systems. Nat. Rev. Microbiol. 1–15.

Marraffini, L.A., and Sontheimer, E.J. (2008). CRISPR Interference Limits Horizontal Gene Transfer in Staphylococci by Targeting DNA. Science (80-.). *322*, 1843–1845.

Mohanraju, P. et al. (2016). Diverse evolutionary roots and mechanistic variations of the CRISPR-Cas systems. Science (80-.). *353*, aad5147.

Mojica, F.J.M. et al. (2009). Short motif sequences determine the targets of the prokaryotic CRISPR defence system. Microbiology *155*, 733–740.

Mulepati, S. et al. (2014). Crystal structure of a CRISPR-RNA guided surveillance compex bound to a ssDNA target. Science (80-.). *345*, 1479–1484.

Nam, K.H. et al. (2012). Cas5d protein processes Pre-crRNA and assembles into a cascade-like interference complex in subtype I-C/Dvulg crispr-cas system. Structure *20*, 1574–1584.

Nuñez, J.K. et al. (2015b). Foreign DNA capture during CRISPR–Cas adaptive immunity. Nature *527*, 535–538.

Nuñez, J.K. et al. (2015a). Integrase-mediated spacer acquisition during CRISPR–Cas adaptive immunity. Nature.

Rao, C. et al. (2016). Active and Adaptive *Legionella* CRISPR-Cas reveals a recurrent challenge to the pathogen. Cell. Microbiol.

Redding, S. et al. (2015). Surveillance and Processing of Foreign DNA by the Escherichia coli CRISPR-Cas System Article Surveillance and Processing of Foreign DNA by the Escherichia coli CRISPR-Cas System. Cell 1–12.

Rollins, M.F. et al. (2015). Mechanism of foreign DNA recognition by a CRISPR RNA-guided surveillance complex from Pseudomonas aeruginosa. Nucleic Acids Res. *43*, 24–25.

Rutkauskas, M. et al. (2015). Directional R-Loop Formation by the CRISPR-Cas Surveillance Complex Cascade Provides Efficient Off-Target Site Rejection. Cell Rep. *10*, 1534–1543.

Sashital, D.G. et al. (2011). An RNA-induced conformational change required for CRISPR RNA cleavage by the endoribonuclease Cse3. Nat. Struct. Mol. Biol. *18*, 680–687.

Sashital, D.G. et al. (2012). Mechanism of Foreign DNA Selection in a Bacterial Adaptive Immune System. Mol. Cell *46*, 606–615.

Semenova, E. et al. (2011). Interference by clustered regularly interspaced short palindromic repeat (CRISPR) RNA is governed by a seed sequence. Proc. Natl. Acad. Sci. U. S. A. *108*, 10098–10103.

Singh, D. et al. (2018). Real-time observation of DNA target interrogation and product release by the RNA-guided endonuclease CRISPR Cpf1 (Cas12a). Proc. Natl. Acad. Sci. *115*, 5444–5449.

Sternberg, S.H. et al. (2015). Conformational control of DNA target cleavage by CRISPR-Cas9. Nature *527*, 110–113.

Szczelkun, M.D. et al. (2014). Direct observation of R-loop formation by single RNA-guided Cas9 and Cascade effector complexes. Proc. Natl. Acad. Sci. U. S. A. *111*, 9798–9803.

van Erp, P.B.G. et al. (2015). Mechanism of CRISPR-RNA guided recognition of DNA targets in *Escherichia coli*. Nucleic Acids Res. gkv793.

Westra, E.R. et al. (2012). CRISPR Immunity Relies on the Consecutive Binding and Degradation of Negatively Supercoiled Invader DNA by Cascade and Cas3. Mol. Cell *46*, 595–605.

Westra, E.R. et al. (2013). Type I-E CRISPR-Cas Systems Discriminate Target from Non-Target DNA through Base Pairing-Independent PAM Recognition. PLoS Genet. *9*.

Wiedenheft, B. et al. (2011). Structures of the RNA-guided surveillance complex from a bacterial immune system. Nature *477*, 486–489.

Xiao, Y. et al. (2017). Structure Basis for Directional R-loop Formation and Substrate Handover Mechanisms in Type I CRISPR- Cas System Article Structure Basis for Directional R-loop Formation and Substrate Handover Mechanisms in Type I CRISPR-Cas System. Cell *170*, 48–60.e11.

Xiao, Y. et al. (2018). Structure basis for RNA-guided DNA degradation by Cascade and Cas3. *0839*, 1–12.

Xue, C. et al. (2015). CRISPR interference and priming varies with individual spacer sequences. Nucleic Acids Res. *43*, 10831–10847.

Xue, C. et al. (2016). Conformational Control of Cascade Interference and Priming Activities in CRISPR Immunity Short Article Conformational Control of Cascade Interference and Priming Activities in CRISPR Immunity. Mol. Cell 1–9.

Xue, C. et al. (2017). Real-Time Observation of Target Search by the CRISPR Surveillance Complex Cascade. Cell Rep. *21*, 3717–3727.

Yosef, I. et al. (2012). Proteins and DNA elements essential for the CRISPR adaptation process in Escherichia coli. Nucleic Acids Res. *40*, 5569–5576.

Zhao, H. et al. (2014). Crystal structure of the RNA-guided immune surveillance Cascade complex in Escherichia coli. Nature.

CHAPTER 5. CONCLUSIONS AND FUTURE DIRECTIONS

Conclusion

Upon infection, bacteria and archaea acquire foreign DNA into the host CRISPR locus as a molecular memory for CRISPR immunity. Cas proteins must select, process, and integrate spacers that can lead to functional target recognition during the interference. The universally conserved Cas1 and Cas2 proteins along with additional adaptation proteins such as Cas4 are required, however the role of Cas4 during this adaptation process remained mysterious. Prior to our work, because Cas4 has exo- or endonuclease activity (Zhang et al., 2012; Lemak et al., 2013, 2014), it has been hypothesized that Cas4 may generate the prespacer substrates for Cas1-Cas2 mediated integration. Using a combination of biochemical and structural analysis, we demonstrated that Bacillus halodurans type I-C Cas4 is responsible for prespacer selection and processing and directly associates with Cas1-Cas2 proteins to form a higher-order complex. We found that Cas4 interacts tightly with Cas1 and the presence of CRISPR DNA substrates stabilizes the complex formation of Cas4-Cas1-Cas2. To determine how Cas4 interacts with the adaptation machinery, we used single-particle electron microscopy (EM) of negatively stained Cas4-Cas1 or Cas4-Cas1-Cas2 complexes. Two Cas4 subunits form a heterohexameric complex with two Cas1 dimers, while one or two Cas4 subunits associate with a symmetrical hetrohexameric Cas1-Cas2 complex, creating symmetrical or asymmetrical Cas4-Cas1-Cas2 complexes. Although Cas4 is an exonuclease, the Cas4-Cas1-Cas2 complex cleaves long 3′ overhangs of prespacers endonucleolytically in PAM dependent manner. In addition, the complex processes directly upstream of the PAMs and integrates the processed prespacers precisely at the leader-repeat junction of the CRISPR locus. Together, these results reveal the central role the Cas4-Cas1-Cas2 complex plays in spacer selection, processing, and integration, ensuring the fidelity of

the CRISPR locus to provide functional spacers for target recognition during interference (Fig. 1).

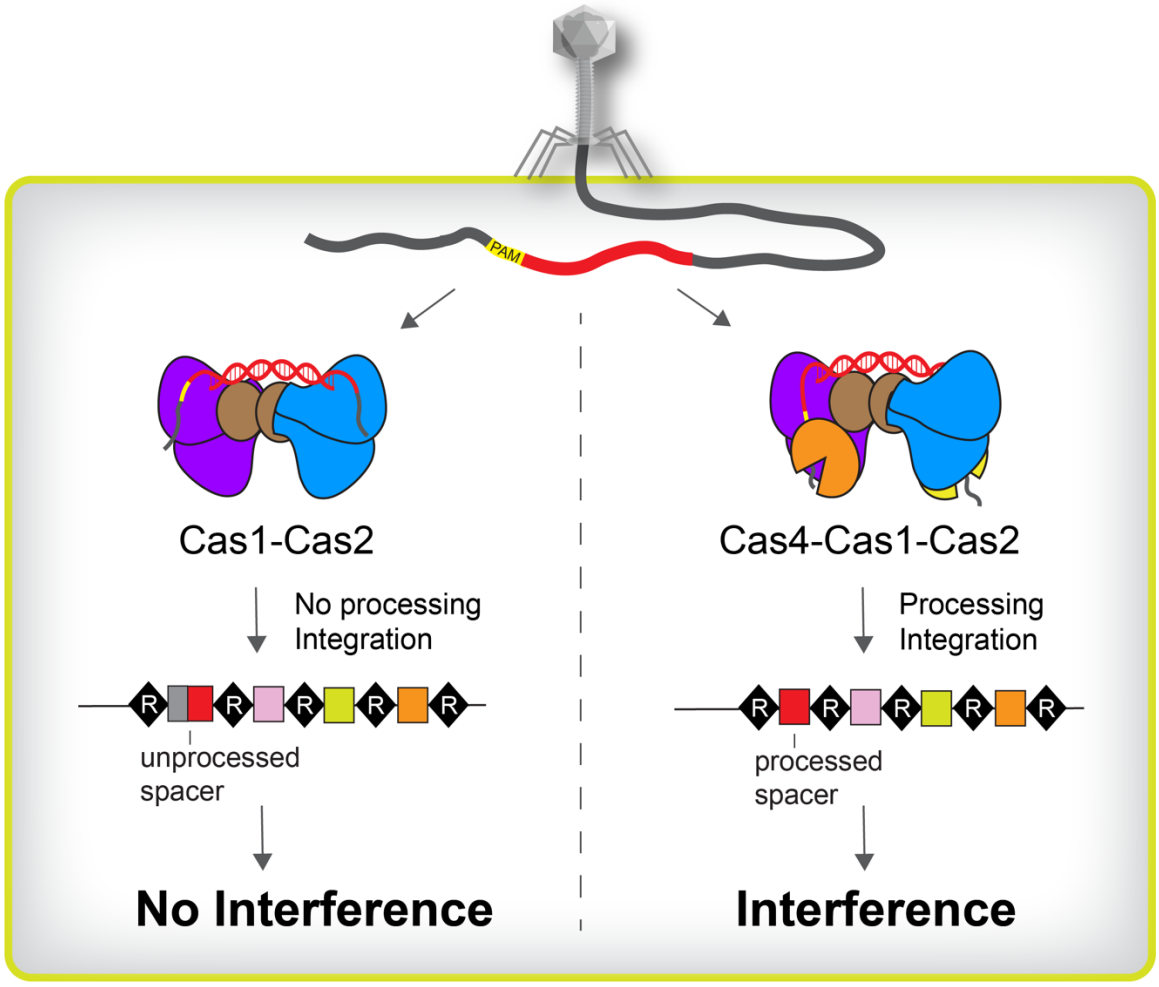

Figure 1. Model of Cas4-Cas1-Cas2 complex during adaptation. Cas1-Cas2 integrates unprocessed spacers into the CRISPR locus, while Cas4-Cas1-Cas2 integrates processed spacers to maintain the fidelity of the locus.

For target recognition, type I-C requires only three proteins to form its surveillance complex. The assembly pathway and molecular architecture of type I-C Cascade has been determined (Hochstrasser et al., 2016), however it is relatively less studied than its counterparts of type I-E and I-F. We investigated binding interactions of type I-C Cascade with DNA and showed that type I-C Cascade exhibits much stronger affinity for non-specific DNA. We hypothesized that the non-specific interactions may be due to longer-lived

associations or one-dimensional sliding contacts with DNA. To test this possibility, we began to construct a smFRET system to directly visualize how type I-C Cascade searches and interacts with DNA targets. Using this system, we observed bulk FRET between Cy3-labelled dsDNA target and Cy5-labelled Cascade, indicating that the two fluorophores are in close proximity and the fluorescent label on Cascade did not affect Cascade DNA binding activity. Overall, we reconstituted type I-C Cascade binding activities in vitro and constructed a system that is suitable for a smFRET assay.

Future Directions

Although considerable progress has been made toward understanding CRISPR adaptation, key aspects of the mechanistic details remain unclear. Our studies showed that *B. halodruans* type I-C Cas4-Cas1-Cas2 complex is highly PAM specific and cleaves directly upstream of the PAM sites. In other type I systems, either the last nucleotide of PAM sequences or additional *cas4* gene partially dictates the orientation during integration. However, most systems lack a second *cas4* gene and type I-C Cas4 removes the entire PAM sequence after processing. Thus, it is unknown how type I-C directionally integrates prespacers into the CRISPR locus to create functional spacers for target recognition. It is also unclear whether Cas4 processes both ends of the prespacer, or only the PAM end. Given the high sequence-specificity of endonucleolytic processing, it is unlikely that the Cas4-Cas1-Cas2 complex processes both ends of the prespacer. Instead, it is possible that Cas4 processes sequence specifically only on PAM-end, while other factor is involved in cleaving at non-PAM sites.

In *E. coli*, the degradation products from RecBCD or Cas3 are used for prespacers by the Cas1-Cas2 adaptation complex. But, most type I systems have Cas4 that exhibits endo- or exonuclease activities. It may be interesting to look at whether Cas4's exo activity may degrade invaders and create a pool of DNA fragments for further processing by Cas4-Cas1-Cas2, which defines the PAM end.

We observed that the type I-C adaptation complex integrates ssDNA substrates into the CRISPR locus. Most type I adaptation complexes are unable to use ssDNA as prespacers or integrate in low efficiency. Typically, the adaptation complex catalyzes integration events via two transesterification reactions mediated by nucleophilic attack of the 3′ hydroxyl on each strand of dsDNA substrates. Although the type I-C adaptation complex integrates ssDNA during half-site integration, it is unclear how ssDNA substrates can lead to full-site integration. However, similar integration events of ssDNA were detected in type III-B system in which reverse transcriptase-Cas1 synthesizes a cDNA from an RNA template for spacer formation. Further work is needed on how the acquired spacers specifically for type I system are used for guiding type III surveillance complex.

Type I-E and type I-F systems are the most studied in understanding target recognition activities by the surveillance complex. Type I-C is the most compact system that can be used for potential biotechnology applications, such as gene repression, but it is relatively understudied. We have investigated the binding interactions of Cascade with DNA and constructed a smFRET system to directly visualize the binding interactions. We speculated that since we observed much stronger non-specific interactions with DNA, Cascade/I-C may dwell longer on DNA in search of PAMs through one-dimensional sliding. Future studies will employ this smFRET system to determine the mechanism and kinetics of how Cascade/I-C searches for DNA targets.

References

Hochstrasser, M.L. et al. (2016). DNA Targeting by a Minimal CRISPR RNA-Guided Cascade. Mol. Cell *63*, 840–851.

Lemak, S. et al. (2013). Toroidal structure and DNA cleavage by the CRISPR-associated [4Fe-4S] cluster containing Cas4 nuclease SSO0001 from sulfolobus solfataricus. J. Am. Chem. Soc. *135*, 17476–17487.

Lemak, S. et al. (2014). The CRISPR-associated Cas4 protein Pcal 0546 from Pyrobaculum calidifontis contains a [2Fe-2S] cluster : crystal structure and nuclease activity. Nucleic Acid Res. *42*, 11144–11155.

Zhang, J. et al. (2012). The CRISPR Associated Protein Cas4 Is a 5′ to 3′ DNA Exonuclease with an Iron-Sulfur Cluster. PLoS One *7*.